

# **Reputation In Everyday Life**

**NICOLE CLAIR ENGELER**

UNIVERSITY COLLEGE LONDON

DEPARTMENT OF EXPERIMENTAL PSYCHOLOGY

Submitted to University College London (UCL) in partial  
fulfilment of the requirements for the award of the degree of  
Doctor of Philosophy.

Primary supervisor: Prof. Nichola J. Raihani

Secondary supervisor: Prof. Daniel C. Richardson

Thesis submission date: 15/07/2024

# **Declaration**

I, Nicole Clair Engeler confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

# Abstract

Humans are a remarkably cooperative species, and it has been proposed that this cooperation evolved in part due to reputational costs and benefits. While existing studies on reputation often rely on controlled simulations and short economic game experiments, my research delves into real-world contexts, exploring how individuals shape and perceive reputations. Using a daily diary phone app, **Chapter 2** investigates how reputation management varies across social relationships and situational contexts in everyday life. Findings were comparable to laboratory and field reports, additionally showing that social relationship factors influence reputational concern more than social visibility cues. **Chapter 3** examines the reputational consequences of punishing in daily life. I used a weekly survey study conducted across six weeks to determine the factors that have the greatest impact on a punisher's reputation. Results revealed that justness and attributed motives of punishing most influenced a punisher's reputation. **Chapter 4** asks how accurate people are at knowing how their cooperative and uncooperative actions influence their reputation. I examined meta-accuracy (whether people know the relative impression they make on others) and meta-bias (whether they systematically under- or over-estimate how favourably they are perceived). In **Chapter 5**, I first used vignettes to investigate the motivations that observers attribute to punishers, and then solicited descriptions of recent punishing behaviours to discover whether observers' motive attributions match punishers' actual motives (observer accuracy) and whether punishers exhibit meta-accuracy. In **Chapter 6**, I present the results of a registered report, using five economic game experiments to determine how deliberating over personal cost and target impact signals trustworthiness in punishment and helping contexts. Collectively, these studies aim to provide insight into how individuals strategically navigate social interactions to manage their reputation across diverse scenarios, with a particular focus on ecologically valid investigations of cooperation and punishment.

# Impact Statement

This thesis presents a comprehensive examination of reputation in real-world contexts, advancing our understanding of human cooperation and punitive behaviours beyond controlled laboratory settings. By employing ecologically valid methods such as daily diaries, weekly survey studies, and vignette experiments, my research addresses gaps in the literature regarding how individuals strategically navigate social interactions to manage their reputation across diverse scenarios. The findings contribute to evolutionary and psychological theories of reputation, highlighting the importance of social relationships, attributed motives, and the accuracy of individuals' perceptions in shaping reputational outcomes.

## Real-World Context and Ecological Validity

One of the primary contributions of this thesis is its emphasis on ecological validity, enhancing the likelihood that my findings can be generalised to everyday life. Most psychology research on reputation relies on artificial settings, but my studies capture behaviours in naturalistic environments, yielding insights that are more applicable to real-world scenarios. By demonstrating that reputation mechanisms observed in controlled environments largely apply to complex, noisy, and naturalistic settings, this research bridges a gap between theory and practice.

## Key Findings and Theoretical Contributions

**Chapter 2** explores reputation management using a daily diary phone app, revealing that social relationship factors influence reputational concern more than social visibility cues. This chapter underscores the significance of context, as well as the importance of examining multiple factors simultaneously.

**Chapter 3** examines the reputational consequences of punishing in daily life through a weekly survey study conducted across six weeks. Findings indicate that the justness and attributed motives of punishing most strongly influence a punisher's reputation, clarifying previous assumptions about the inherent reputational benefits of third-party punishment.

**Chapter 4** investigates how accurately individuals perceive the impact of their actions on their reputation. The study reveals that actors accurately anticipate perceptions of their cooperative and uncooperative actions, whilst still displaying biases in meta-perceptions.

**Chapter 5** uses vignettes and real-world descriptions to explore punishers' motives and perception accuracy. Observers correctly linked third-party punishment to prosocial motives and second-party punishment to self-centred concerns. Punishers accurately anticipated these views. Despite this, biases persisted; for example, observers exhibited a systematic negativity bias, being suspicious of punishers' motives.

**Chapter 6** extends research on helping behaviours, finding that uncalculated helping signals trustworthiness, whereas uncalculated punishment does not reliably do so. This highlights the context-dependent nature of reputational strategies and the complexity of punitive behaviours.

## **Practical Implications**

These findings can have implications for individuals, organisations, and policymakers. Understanding that reputational concern is heightened by the value placed on relationships can inform strategies to enhance workplace dynamics and promote cooperative environments. Additionally, understanding that our motives may not be as transparent to others as we believe may encourage more effective communication.

## **Conclusion**

This thesis advances our understanding of reputation in everyday life, emphasising the need for ecologically valid research to capture the complexity of human social interactions. By highlighting the importance of situational context, social relationships, and motive attributions, this research provides valuable insights for enhancing cooperation and addressing challenges in various real-world settings.

# UCL Research Paper Declaration Form

## Referencing the doctoral candidate's own published work(s)

**1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)

**a) What is the title of the manuscript?**

Acting Without Considering Personal Costs Signals Trustworthiness in Helpers but Not Punishers

**b) Please include a link to or doi for the work**

<https://doi.org/10.1038/s44271-024-00092-7>

**c) Where was the work published?**

Communications Psychology

**d) Who published the work? (e.g. OUP)**

Springer Nature Limited

**e) When was the work published?**

24<sup>th</sup> May 2024

**f) List the manuscript's authors in the order they appear on the publication**

Nicole C. Engeler, Nichola J. Raihani

**g) Was the work peer reviewed?**

Yes

**h) Have you retained the copyright?**

Yes

**i) Was an earlier form of the manuscript uploaded to a preprint server? (e.g. medRxiv). If 'Yes', please give a link or doi)**

The Stage 1 manuscript of this registered report was published at  
<https://doi.org/10.6084/m9.figshare.24559462.v1>.

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

**2. For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

**a) What is the current title of the manuscript?**

Click or tap here to enter text.

**b) Has the manuscript been uploaded to a preprint server? (e.g. medRxiv; if 'Yes', please give a link or doi)**

Click or tap here to enter text.

c) **Where is the work intended to be published?** (e.g. journal names)

Click or tap here to enter text.

d) **List the manuscript's authors in the intended authorship order**

Click or tap here to enter text.

e) **Stage of publication** (e.g. in submission)

Click or tap here to enter text.

**3. For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

N.E. and N.R. developed the study concept. N.E. designed the study with revisions from N.R. and conducted data collection and analysis; N.E. drafted the initial manuscript; N.E. and N.R. revised and reviewed the manuscript and approved the final manuscript for submission.

**4. In which chapter(s) of your thesis can this material be found?**

Chapter 6. Deliberating Cost and Impact: Trustworthiness Signals in Punishment and Helping

**5. e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

*Candidate*

Nicole C. Engeler

*Date:*

10<sup>th</sup> June 2024

*Supervisor/ Senior Author (where appropriate)*

Nichola Raihani

*Date*

**21/6/24**

# Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor, **Nichola** Raihani. Nichola's guidance has profoundly influenced not only my research but also my overall comprehension of academia and the scientific process. Her knowledge has been invaluable, and I feel fortunate to have benefited from her leadership and support, no matter the continent or time zone.

I am also immensely grateful to **Mark** Dyble, whose empathy, perspective, and willingness to engage with my experience sampling studies have been invaluable. A heartfelt thanks to **Jim** Allen for sharing his insights and providing support as a member of my thesis committee. His readiness to assist whenever I was stuck is greatly appreciated, and provided me with reassurance. I also extend my gratitude to **Mirco** Musolesi, my committee chair, for being an enthusiastic and reliable presence throughout this journey.

I would like to express my appreciation to the undergraduate students who took an interest in my work and helped collect participants for some of my pilot studies. A heartfelt thank you also to **Henrik** Singmann for his helpful advice on Bayesian methods. Thank you to **Jeff** Lees for his advice on the use of his Social Accuracy Model adaptation.

I am fortunate to have been part of the cross-disciplinary Ecological Brain DTP, which places great value on understanding how humans operate in the real world. Many thanks therefore to the Leverhulme Trust for making this PhD possible. I am grateful to my EcoBrain peers, and to Warda for always being so kind, efficient, and a wonderful presence. Special thanks to **Gabriella** Vigliocco, for sharing her enthusiasm for multidisciplinary research and for always making it clear that she is in my corner. Thanks to **Daniel** Richardson, **Nick** Tyler, and **Sam** Gilbert for providing incredible rotation projects in their labs, where I gained a wealth of knowledge and had a great time, despite some disruptions due to COVID-19 (and realising that I am not a fan of the opera). Additionally, thank you, Daniel, for agreeing to be my second supervisor later on.

I am deeply grateful to my family and friends, who have been so supportive; especially those who were happy to act as guinea pigs for my various studies. Thank you to my sister **Kitty** for always listening to me and encouraging me with her unwavering belief in me. Thank you to **Jula**, who first showed me how beautiful R figures can be, and has been a true friend no matter how many countries lie between us. **Joanna**, our chats have been invaluable, and I am so



grateful for your support and confidence in me throughout this journey; your food pictures also always kept me going. Thank you to our other 'Naked Chimps', **Jamie** and **Kelsey**, for staying in supportive contact. Thank you to **Mette** for finding ways to support me even when I said I had no time, and for always looking out for me and making plans we can look forward to. Thank you to **Krissi** for being in the same boat and understanding the journey. Thank you, **Julie**, for consistently staying in touch and being there every step of the way; no matter our stages in life you care to see me succeed. Thank you to **Rere** for staying engaged and asking for updates on my PhD journey. **Fredi**, **Tim** and **Mattis** – thank you for being so lovely and helping me stay focused yet sane during the last few months of this PhD. And to **Rebecca**, thank you for being an amazing and supportive neighbour. Your generosity in letting me stay in your home when my household had COVID and your kindness in making me meals were much appreciated.

A special thank you to my fiancé **Harry** for being there for me 'in sickness and in health,' for proofreading numerous chapters, for being understanding of my odd work requirements, and for believing that I could achieve this. Thank you for celebrating my successes, comforting me during setbacks, and always reminding me of my potential.

Above all, I thank my parents, **Ralph** and **Joanne**, who have gone above and beyond in supporting me during this PhD. They have taken on anything they possibly could to lighten my load, allowing me to focus on my work. Their boundless generosity, practical assistance, and unwavering love and support have been invaluable to me.

# Table of Contents

|                                                                                            |           |
|--------------------------------------------------------------------------------------------|-----------|
| Abstract.....                                                                              | 3         |
| Impact Statement.....                                                                      | 4         |
| UCL Research Paper Declaration Form.....                                                   | 6         |
| Acknowledgements.....                                                                      | 8         |
| Table of Contents.....                                                                     | 10        |
| <b>1 General Introduction.....</b>                                                         | <b>15</b> |
| 1.1 Cooperation and Reputation.....                                                        | 16        |
| 1.1.1 Reputation Mechanisms: Indirect Reciprocity and Reputation-Based Partner Choice..... | 18        |
| 1.1.2 Interplay of Reputation Management and Impression Formation.....                     | 21        |
| 1.1.3 The Cognitive Mechanisms Supporting Reputation-Based Cooperation.....                | 23        |
| 1.1.4 A Developmental Approach to Impression Formation and Management.....                 | 25        |
| 1.2 Punishment and Reputation.....                                                         | 27        |
| 1.3 Relevant Methodological Frameworks.....                                                | 34        |
| 1.4 Overview of this Thesis.....                                                           | 39        |
| <b>2 Reputation Management In Everyday Life.....</b>                                       | <b>42</b> |
| 2.1 Introduction.....                                                                      | 43        |
| 2.2 Methods.....                                                                           | 48        |
| 2.2.1 Participants.....                                                                    | 48        |
| 2.2.2 Intake Questionnaire.....                                                            | 48        |
| 2.2.3 Daily Assessment Phase.....                                                          | 49        |
| 2.2.4 Survey Content.....                                                                  | 50        |
| 2.2.5 Participant Payment.....                                                             | 52        |
| 2.2.6 Analysis Plan.....                                                                   | 52        |
| 2.3 Results.....                                                                           | 55        |

|          |                                                                                                                       |            |
|----------|-----------------------------------------------------------------------------------------------------------------------|------------|
| 2.3.1    | Reputational Concern Across Reputation Management and Social Interaction Events.....                                  | 55         |
| 2.3.2    | Preregistered Analysis: Individual Factors Predicting Levels of Reputational Concern.....                             | 56         |
| 2.3.3    | Additional Holistic Analysis: Combined Predictors of Reputational Concern.....                                        | 60         |
| 2.3.4    | The Influence of Traits on Concern for Reputation.....                                                                | 61         |
| 2.4      | Discussion.....                                                                                                       | 63         |
| <b>3</b> | <b>The Reputation of Punishers in Daily Life .....</b>                                                                | <b>69</b>  |
| 3.1      | Introduction.....                                                                                                     | 70         |
| 3.2      | Methods.....                                                                                                          | 75         |
| 3.2.1    | Participants.....                                                                                                     | 75         |
| 3.2.2    | Weekly Surveys.....                                                                                                   | 75         |
| 3.2.3    | Survey Content.....                                                                                                   | 76         |
| 3.2.4    | Analysis Plan.....                                                                                                    | 79         |
| 3.3      | Results.....                                                                                                          | 83         |
| 3.3.1    | Overall Reflection of Punishment on the Punisher.....                                                                 | 86         |
| 3.3.2    | The Influence of Contextual Factors on Punishers' Prosocial Reputations .....                                         | 86         |
| 3.3.3    | The Influence of Contextual Factors on Punishers' Formidability Reputations.....                                      | 90         |
| 3.3.4    | The Influence of Contextual Factors on Observers' Behavioural Intentions Towards Punishers.....                       | 90         |
| 3.3.5    | An Overview of the Contextual Factors Influencing Punishers' Reputations.....                                         | 91         |
| 3.3.6    | Relationship Effects on Punishers' Formidable versus Prosocial Reputations.....                                       | 93         |
| 3.4      | Discussion.....                                                                                                       | 95         |
| <b>4</b> | <b>Meta-Cognitive Accuracy in Assessing the Influence of Cooperative and Uncooperative Actions on Reputation.....</b> | <b>101</b> |
| 4.1      | Introduction.....                                                                                                     | 102        |
| 4.2      | Methods.....                                                                                                          | 108        |
| 4.2.1    | Sampling Plan.....                                                                                                    | 108        |
| 4.2.2    | Participants and Design.....                                                                                          | 108        |

|          |                                                                                                                            |            |
|----------|----------------------------------------------------------------------------------------------------------------------------|------------|
| 4.2.3    | Procedure.....                                                                                                             | 109        |
| 4.2.4    | Analysis Plan.....                                                                                                         | 112        |
| 4.3      | Results.....                                                                                                               | 118        |
| 4.3.1    | Actor Meta-Accuracy.....                                                                                                   | 118        |
| 4.3.2    | Moderators of Meta-Accuracy.....                                                                                           | 123        |
| 4.3.3    | Meta-Bias.....                                                                                                             | 126        |
| 4.3.4    | Observer Accuracy.....                                                                                                     | 129        |
| 4.3.5    | Secondary Analyses .....                                                                                                   | 134        |
| 4.4      | Discussion.....                                                                                                            | 138        |
| <b>5</b> | <b>Understanding Punisher Motives and Reputation: Meta-Perceptions and Observer Inferences Surrounding Punishment.....</b> | <b>144</b> |
| 5.1      | Investigating Observer Perceptions of Motives in Second- and Third-Party Punishers through Vignettes.....                  | 145        |
| 5.1.1    | Introduction.....                                                                                                          | 145        |
| 5.1.2    | Methods.....                                                                                                               | 149        |
| 5.1.2.1  | Participants.....                                                                                                          | 149        |
| 5.1.2.2  | Procedure and Materials.....                                                                                               | 149        |
| 5.1.2.3  | Analysis Plan.....                                                                                                         | 151        |
| 5.1.3    | Results.....                                                                                                               | 153        |
| 5.1.4    | Discussion.....                                                                                                            | 158        |
| 5.2      | Meta-Cognitive and Observer Accuracy Surrounding Punisher Motive Attributions.....                                         | 162        |
| 5.2.1    | Introduction.....                                                                                                          | 162        |
| 5.2.2    | Methods.....                                                                                                               | 167        |
| 5.2.2.1  | Sampling Plan.....                                                                                                         | 167        |
| 5.2.2.2  | Participants.....                                                                                                          | 167        |
| 5.2.2.3  | Procedure.....                                                                                                             | 168        |
| 5.2.2.4  | Analysis Plan.....                                                                                                         | 172        |
| 5.2.3    | Results.....                                                                                                               | 174        |
| 5.2.3.1  | Actor Meta-Accuracy.....                                                                                                   | 174        |
| 5.2.3.2  | Moderators of Meta-Accuracy.....                                                                                           | 179        |
| 5.2.3.3  | Meta-Bias.....                                                                                                             | 180        |
| 5.2.3.4  | Observer Accuracy.....                                                                                                     | 182        |
| 5.2.3.5  | Moderators of Observer Accuracy.....                                                                                       | 186        |

|                                                                                                                                                               |            |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| 5.2.3.6 Observer Bias.....                                                                                                                                    | 187        |
| 5.2.3.7 Item Relevance for SPP versus TPP.....                                                                                                                | 189        |
| 5.2.3.8 Political Orientations and Punishing Sentiments.....                                                                                                  | 196        |
| 5.2.4 Discussion.....                                                                                                                                         | 197        |
| <b>6 Deliberating Cost and Impact: Trustworthiness Signals in Punishment and Helping.....</b>                                                                 | <b>205</b> |
| 6.1 Introduction.....                                                                                                                                         | 206        |
| 6.2 Methods.....                                                                                                                                              | 215        |
| 6.2.1 Ethics information.....                                                                                                                                 | 215        |
| 6.2.2 Design.....                                                                                                                                             | 215        |
| 6.2.3 Power Analysis.....                                                                                                                                     | 219        |
| 6.2.4 Participants.....                                                                                                                                       | 220        |
| 6.2.5 Data Exclusion.....                                                                                                                                     | 220        |
| 6.3 Analysis Plan.....                                                                                                                                        | 221        |
| 6.3.1 Bayesian Analyses.....                                                                                                                                  | 222        |
| 6.4 Deviations from the Preregistration.....                                                                                                                  | 224        |
| 6.5 Results.....                                                                                                                                              | 225        |
| 6.5.1 Is deliberation around help and punishment used as a signal of trustworthiness?.....                                                                    | 225        |
| 6.5.2 Is deliberation around help and punishment <i>perceived</i> as a signal of trustworthiness?.....                                                        | 227        |
| 6.5.3 Is deliberation around help and punishment <i>actually</i> a signal of trustworthiness?.....                                                            | 229        |
| 6.5.4 Do trust and trustworthiness differ across experiments?.....                                                                                            | 231        |
| 6.5.5 Is deliberation around decisions <i>not</i> to help or punish perceived as a signal of trustworthiness?.....                                            | 233        |
| 6.5.6 Is deliberation around decisions <i>not</i> to help or punish a signal of trustworthiness?.....                                                         | 233        |
| 6.5.7 Does deliberation have a stronger influence on perceived and actual trustworthiness when decisions result in helping or punishing versus inaction?..... | 234        |
| 6.5.8 Additional Unregistered Analyses: Does deliberation have a differing influence on trust and trustworthiness for non-helpers versus non-punishers?.....  | 236        |
| 6.6 Discussion.....                                                                                                                                           | 238        |
| <b>7 General Discussion.....</b>                                                                                                                              | <b>243</b> |
| 7.1 Summary of Context and Rationale.....                                                                                                                     | 244        |
| 7.2 Summary of Findings.....                                                                                                                                  | 244        |
| 7.3 General Implications.....                                                                                                                                 | 246        |
| 7.3.1 Advancements in Research and Theory.....                                                                                                                | 246        |
| 7.3.2 Practical Implications and Applications.....                                                                                                            | 252        |
| 7.4 General Limitations.....                                                                                                                                  | 253        |

|       |                                                                                            |     |
|-------|--------------------------------------------------------------------------------------------|-----|
| 7.4.1 | Experience Sampling Methodologies.....                                                     | 254 |
| 7.4.2 | WEIRD Participants and Socio-Ecological Factors.....                                       | 256 |
| 7.4.3 | Limited Investigations of Indirect Reciprocity and Reputation-Based<br>Partner Choice..... | 257 |
| 7.4.4 | Other Limitations of Scope.....                                                            | 258 |
| 7.5   | Future Studies.....                                                                        | 259 |
| 7.5.1 | Metacognition and Reputation.....                                                          | 259 |
| 7.5.2 | Making Use of Other Games: Cyberball.....                                                  | 260 |
| 7.5.3 | Reputation Management and (Avoidance of) Embarrassment.....                                | 261 |
| 7.5.4 | Reputation and Individual Differences.....                                                 | 262 |
| 7.5.5 | Research Linked to Open-Text Data.....                                                     | 263 |
| 7.6   | Conclusion.....                                                                            | 263 |
|       | Appendices.....                                                                            | 265 |
|       | Appendix A: Supplementary Information for Chapter 2.....                                   | 266 |
|       | Appendix B: Supplementary Information for Chapter 3.....                                   | 290 |
|       | Appendix C: Supplementary Information for Chapter 4.....                                   | 327 |
|       | Appendix D: Supplementary Information for Chapter 5.....                                   | 373 |
|       | Appendix E: Supplementary Information for Chapter 6.....                                   | 443 |
|       | Bibliography.....                                                                          | 513 |

# Chapter 1

## General Introduction

*"Your devotion is nothing more than cowardice. You would not be here if you had anywhere else to go."*

— **Lord Voldemort, Harry Potter and the Goblet of Fire**

Voldemort's accurate assessment of Wormtail's devotion as cowardice highlights the complexities inherent in reputation management and observer perceptions. His words emphasise the importance of discerning genuine intentions from self-interested motives when evaluating reputations and making informed decisions about social interactions.

This introduction sets the stage for an exploration into the multifaceted dimensions of reputation, delving into its role in discerning trustworthy social partners – particularly in the realm of cooperative and punitive behaviours. While existing studies in psychology often rely on controlled simulations and economic games, my research adopts more ecologically valid methodologies to understand reputation in real-life contexts. Using approaches such as a weekly survey study, a daily diary phone app, and vignettes, this thesis explores reputations across diverse contexts. Central to this investigation is the examination of factors influencing reputation concern, the accuracy of individuals' perceptions regarding the impact of their cooperative and punitive actions on their reputation, and the attribution of motives to observed behaviours. Additionally, this research delves into the factors influencing reputational consequences of punishment and how deliberating over personal cost and target impact signals trustworthiness in punishment and helping contexts. Overall, the aim of the present research is to deepen our understanding of how individuals strategically navigate social interactions to cultivate their reputations and identify reliable interaction partners based on their reputation. By doing so, this thesis contributes to a richer understanding of human behaviour in social settings.

## **1.1 Cooperation and Reputation**

Cooperation, an interaction outcome that yields fitness benefits (typically defined as increases in survival and reproductive success; see Orr, 2009 for a review of the various definitions of fitness) to all parties (Bshary & Bergmüller, 2008), is regarded as one of humanity's most remarkable attributes. Our species is unique in its large-scale, non-kin cooperation: we cooperate with non-relatives with whom we are unlikely to interact in future, where benefits of direct reciprocity (a mechanism where people help those who have helped them in the past) are



low (Boyd, 2017; Bshary & Raihani, 2017; Gächter & Herrmann, 2009; Henrich, 2017; Henrich et al., 2005; Melis & Semmann, 2010; Raihani & Bshary, 2015a; Rand & Nowak, 2013). Indeed, our propensity to cooperate even with individuals who are not biologically related is considered a key human characteristic (Granovetter, 1978; Smith & Szathmáry, 1995; Noë & Hammerstein, 1995; Kollock, 1998). In daily life people for instance give up their seats so that elderly or pregnant people can sit, hold the elevator for others, provide help to someone asking for directions, give money to a homeless person, or donate to aid people struck by disaster on another continent.

Although cooperation yields collective benefits, it requires at least one individual to invest in the interaction by incurring a temporary cost (e.g. by expending energy or time, increasing risks of harm, or forgoing other profitable opportunities), either to help another or to contribute to a public good, i.e., a shared resource or benefit (Melis & Raihani, 2023). Cooperation has therefore been considered a social dilemma and evolutionary challenge (Fehr & Fischbacher, 2003; Rand & Nowak, 2013; Pennisi, 2009). Social dilemmas, which occur when collective interests conflict with individual interests, have been extensively modelled using public goods games (PGGs). In the standard design, four players receive an endowment from the experimenter. Players can keep the endowment or contribute all or part of it to a common pool. The amount in the pool is multiplied by a factor larger than one and smaller than the number of players, and then shared equally among all players, regardless of their contributions (Ledyard, 1995). In PGGs, individuals are tempted to keep their endowment and enjoy the rewards of others' contributions to the shared resource. However, if everyone contributed their entire endowment, the group would be better off. The possibility to benefit from others' cooperation without contributing oneself presents the dilemma of how large-scale non-kin cooperation can arise and be sustained.

Reputation has been proposed as one key solution to the scale of cooperation found in humans (Alexander, 1979, 1987; Barclay, 2012; Nowak & Sigmund, 1998, 2005; Suzuki & Akiyama, 2005; Wu et al., 2016a, 2016b). Generally, reputation refers to (shared) assessments of an individual's qualities, which includes cooperativeness (Giardini & Wittek, 2019; Tennie et al., 2010). These assessments may stem from familiarity with another's prior actions, obtained through past interactions, direct observation, or the sharing of information about others, commonly known as gossip (Barkow, 1992; Does Cruz et al., 2021; Emler, 2019). These reputations are subjective and can be inaccurate, as well as be subject to variation due to misperceptions, biases, or differing interaction histories among individuals. Reputations are

also trait-specific, reflecting others' beliefs about whether an individual possesses certain traits (Barclay, 2015). These traits can span physical attributes (such as strength), dispositional characteristics (like trustworthiness, reliability, or willingness to engage in conflict), social standing (such as having influential allies), or a combination thereof (Barclay, 2015). While there may be overlap between traits, with reputations in one domain influencing perceptions in related areas, an overall "good reputation" suggests that individuals hold positive views of a person across multiple relevant traits (Barclay, 2015).

### **1.1.1 Reputation Mechanisms: Indirect Reciprocity and Reputation-Based Partner Choice**

Obtaining a good reputation can lead to downstream benefits through two primary routes: individuals may be more likely to be chosen as a partner (reputation-based partner choice, (Roberts, 1998)) or they may be more likely to be rewarded by others (indirect reciprocity; Alexander, 1987; Boyd & Richerson, 1989; Kandori, 1992; Ohtsuki & Iwasa, 2007) (see Roberts et al., 2021 for an in-depth comparison of the two).

Reputation-based theories of cooperation propose that individuals observe and assess the actions of others, and adjust their own behaviour accordingly (Barclay, 2016; Nowak & Sigmund, 2005; Roberts, 1998). One prominent theory is indirect reciprocity (IR; Alexander, 1987). This theory posits that when someone incurs a short-term cost ( $c$ ) to benefit another person, they become more likely to receive reciprocal benefits ( $b$ ) from observers, resulting in a net gain when the benefit ( $b$ ) outweighs the cost ( $c$ ) (Boyd & Richerson, 1989). Specifically, in frameworks describing IR (Alexander, 1987; Boyd & Richerson, 1989; Nowak & Sigmund, 1998a; Panchanathan & Boyd, 2004), individuals make decisions to cooperate or not, and these decisions contribute to an 'image score' that influences whether uninvolved third parties will cooperate with those individuals in future interactions (Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998a). An image score represents a (public) record of previous actions and is revised following each interaction. According to Nowak and Sigmund's original model (1998), people can then use an individual's image score as a basis for determining whether to provide assistance. Indeed, individuals cooperate more when presented with an incentive to gain reputation via IR (Milinski et al., 2002, 2006; Rockenbach & Milinski, 2006; Semmann et al., 2004, 2005) and people give preferentially to individuals with positive image scores (Milinski et al., 2001; Seinen & Schram, 2006; Semmann et al., 2004; Wedekind & Milinski, 2000).

However, determining how an individual's actions ought to be judged is not straightforward. Various approaches for bestowing reputations based on actors' previous behaviours have been proposed and examined in the literature (e.g., Gross & De Dreu, 2019; Milinski et al., 2001; Ohtsuki & Iwasa, 2004, 2006; Pacheco et al., 2006; Panchanathan & Boyd, 2003; Santos et al., 2018). Not all of these methods are effective in promoting cooperation. Indirect reciprocity is only evolutionarily stable when group members have a consensus on when individuals are expected to cooperate and when it is acceptable to defect (i.e., fail to cooperate). Such 'rules' rely on social norms (Nowak & Sigmund, 2005; Ohtsuki & Iwasa, 2004) and involve moral judgements (Alexander, 1987; Efferson & Fehr, 2018; Sigmund, 2012). For instance, if a norm permits withholding cooperation from a defector, individuals can choose not to cooperate without risking damage to their reputation (Roberts et al., 2021). Among the consensus-based evaluations, only eight norms, known as the "leading eight," have been identified as capable of sustaining cooperation while also being resilient to observational errors and mutation (Ohtsuki & Iwasa, 2004, 2006). These leading eight norms share several joint properties: they reward cooperation among individuals with good reputations, penalise defection against individuals with good reputations, maintain the good reputation of actors who defect against those with bad reputations, and extend forgiveness by assigning good reputation to actors with bad reputations if they cooperate with individuals with good reputations (Ohtsuki & Iwasa, 2004, 2006).

Besides IR, individuals may also benefit from a cooperative reputation via reputation-based partner choice (RBPC), where a good reputation makes one more likely to be chosen as a partner for future interactions. Having a reputation for being cooperative can act as a signal, where signals are considered as observable characteristics adapted to influence the actions of a receiver in a manner advantageous to the signaller (Smith & Harper, 2003). Individuals benefit from selecting partners who exhibit high quality and good intentions, implying they are capable and willing to provide benefits (Barclay, 2013). The hypothesis of assortative pairing (originating from models examining the correlation between generosity and selectiveness; McNamara et al., 2008; Sherratt & Roberts, 1998) suggests that those perceived as most cooperative should form partnerships with similarly helpful individuals (Lyon & Montgomerie, 2012; Nesse, 2007; West-Eberhard, 1983). After forming these partnerships, individuals establish a cooperative relationship that can involve direct reciprocity, incidental benefits, or other mechanisms of interdependence between partners. Thus, the signaller gains from their effort in signalling, while the recipient benefits from exercising selectivity in choosing partners. As being seen as cooperative enhances the likelihood of being selected for future interactions, the signal perceiver is not necessarily required to make significant investments to repay the

signaller for their cooperative actions. Rather, being selected for a mutually beneficial interaction could serve as compensation for the cooperator's initial investment (Bshary & Grutter, 2006; McNamara et al., 2008).

As partner choice enables cooperative individuals to seek out interaction partners with favourable reputations (e.g., Barclay & Willer, 2007; Sylwester & Roberts, 2010, 2013) whilst avoiding those with negative reputations (Rockenbach and Milinski, 2011), it can foster competition for potential cooperative partners in what are termed biological markets (Barclay, 2013, 2016; Barclay & Willer, 2007; Noë & Hammerstein, 1994; Roberts, 1998). This is also often referred to as “competitive altruism” as competition to be chosen as social partner can lead to an escalation of prosocial behaviour (Barclay, 2004, 2010, 2011, 2012, 2013, 2016; Barclay & Willer, 2007; Böhm & Regner, 2013; Hardy & Van Vugt, 2006; Herrmann et al., 2019; Macfarlan et al., 2012; Raihani & Smith, 2015; Roberts, 1998; Sylwester & Roberts, 2010, 2013; Van Vugt et al., 2007). Thus, in scenarios where individuals vie for access to partners, RBPC theory suggests that the exhibition of progressively more costly behaviours serve as signals to secure good partners: cooperative actions become “competitive” because the crucial factor lies in the extent of assistance provided relative to others (Barclay, 2004, 2013; Roberts, 1998; Van Vugt et al., 2007). Those who produce the strongest signals for being helpful are then preferentially chosen as partners (Barclay, 2011; Lotem et al., 2003; Roberts, 1998). If reputation provides a reliable and clear indication of an individual's cooperativeness, defectors are at a disadvantage compared to cooperators. This is because cooperators can interact exclusively with each other, thereby reaping the benefits of mutual cooperation. Indeed, in a model in which individuals interacted within their immediate network, severed connections with those who had a bad reputation and formed new bonds with those possessing the highest reputation (or with a randomly selected agent), defectors were deserted, leading to the emergence of stable and elevated levels of cooperation (Fu et al., 2008).

Evidence from real-world contexts (Hardy & Norgaard, 2016; Milinski et al., 2002a; Przepiorka et al., 2017; Raihani & Smith, 2015; Resnick et al., 2006; Yoeli et al., 2013) across diverse societies (Bliege Bird et al., 2012; Bliege Bird & Power, 2015; Gurven et al., 2000; Lyle & Smith, 2014; Macfarlan et al., 2012, 2013; Macfarlan & Lyle, 2015; Power & Ready, 2018; Smith et al., 2003; Smith & Bird, 2000; von Rueden et al., 2019) underscores the advantages associated with having a good reputation, both through IR and RBPC. Individuals with a prosocial reputation are more likely to receive help (Bliege Bird & Power, 2015; Gurven et al., 2000; Lyle & Smith, 2014; Macfarlan et al., 2012, 2013; Power & Ready, 2018) and have an

increased number of social partners (and thereby resources) (Garfield et al., 2021). For instance, Martu hunters from Australia who showed greater generosity in distributing foraged meat had a higher number of cooperative associates and occupied a more prominent position within the social network (Bliege Bird & Power, 2015). Further, in Andean communities, individuals with a cooperative reputation had a larger number of social support connections (Lyle & Smith, 2014), and in Dominican villages, men with reputations for being prosocial could assemble larger labour groups and cultivate more reciprocal partnerships (Macfarlan et al., 2012). In Australia, Meriam men who contributed by hunting turtles (large game which is shared) gained social recognition, as well as more and higher quality mates (Smith et al., 2003).

### **1.1.2 Interplay of Reputation Management and Impression Formation**

In sum, those with a good reputation are often rewarded or sought out, while those with a bad reputation can be condemned and avoided, even ostracised (Milinski et al., 2002b; Over, 2016) (note, however, that while most models presented in this introduction use a simplified system of ‘good’ compared to ‘bad’ reputation, reputation systems in real-life are more complex; Barclay & Barker, 2020; Bartley & Child, 2014; Dafoe et al., 2014; Dukerich & Carter, 2000; Fombrun & Rindova, 1996; Heugens et al., 2004; Macfarlan et al., 2013; Rhee & Haunschild, 2006; Rogers et al., 2018; Tadelis, 2016). People are conscious of this and therefore engage in strategic actions to influence how others perceive them – in other words, manage their reputation. Influencing how others perceive us is fundamental in social interactions (Baumeister, 1982; Goffman, 1958). Reputation management involves strategic behaviours aimed at controlling how others judge us, driven by concerns of how one is viewed by other individuals (Banerjee, 2002). We often act with the aim of eliciting specific (usually positive) evaluations of our reputation from those around us (Goffman, 1958). For instance, people frequently behave more prosocially when their behaviour can be observed and their reputations are therefore under scrutiny (Andreoni & Petrie, 2004; Barclay, 2013; Bateson et al., 2006; Bereczkei et al., 2007; Bradley et al., 2018; Engelmann & Zeller, 2016; Harbaugh, 1998; Hoffman et al., 1994; Milinski et al., 2002a, 2002b; Suzuki & Akiyama, 2005; Sylwester & Roberts, 2010; Van Vugt et al., 2007; Yoeli et al., 2013). Increased visibility can promote cooperation across various areas, such as adhering to tax regulations (Coricelli et al., 2010), participating in elections (Gerber et al., 2008), donating blood (Lacetera & Macis, 2010), conserving energy (Yoeli et al., 2013), and environmentalism (Barclay & Barker, 2020).

Conversely, people are aware that others attempt to manage their own reputations. To form accurate impressions of others, we must therefore infer the intentions behind their actions, which can be challenging as individuals may actively seek to shape how they are perceived. When engaging in reputation management, individuals must therefore acknowledge that others will endeavour to discern the motives behind their actions and strive to predict the impressions they will form of them. Both the ability to evaluate the reputation of others and manage one's own reputation rely on "epistemic vigilance", a set of cognitive abilities likely honed by evolution to aid in distinguishing trustworthy information from potential misinformation (discussed more in the next section) (Sperber et al., 2010). When individuals perceive a vested interest in manipulating their beliefs or influencing their behaviour, they tend to scrutinise actions or information more critically, potentially discounting them. For this reason, cooperative behaviours are not always evaluated positively (reviewed in Berman & Silver, 2022; Raihani & Power, 2021), and can even be punished (Bahry & Wilson, 2006; Irwin & Horne, 2013; Kawamura & Kusumi, 2020). Cooperative individuals frequently encounter scepticism, occasionally being portrayed as insincere self-promoters, superficial virtue-signallers, or sanctimonious hypocrites. Generous individuals are more prone to facing social repercussions when observers perceive their benevolent actions as driven by strategic or self-serving motives (or when observers interpret their helpful behaviour as detrimental to their own interests; Raihani & Power, 2021). Observers may for instance be suspicious regarding the genuine nature of help when the helper appears to deliberate or weigh the costs before extending aid (Hoffman et al., 2015). Conversely, individuals who cooperate without prior knowledge of the associated costs (Jordan et al., 2016) or make swift decisions (Cricher et al., 2013; Evans & van de Calseyde, 2017; Jordan et al., 2016) are often perceived as more genuinely prosocial.

As actors are aware that their prosocial actions could be attributed to ulterior motives, they have devised strategies to communicate their generosity while minimising potential damage to their reputation. For instance, time donations, considered more morally commendable than monetary contributions, often take precedence, reflecting a preference for actions perceived as personally taxing over those with potentially greater impact (Johnson & Park, 2021). Individuals donating to charity may also opt for tasks perceived as arduous or challenging to diminish any suspicions of self-interest underlying ostensibly cooperative behaviour (Olivola & Shafir, 2013). Furthermore, Martu hunters sometimes engage in "pecuniary distancing" tactics, deliberately relinquishing control over the distribution of their catch to forestall inferences of selfish motives from others (Bliege Bird & Power, 2015). Actors can also focus on more subtle displays of generosity, revealed to only a select few, to avoid suspicion of a broader reputation-enhancing

agenda (Bliege Bird et al., 2018; Hoffman et al., 2018). Alternatively, individuals seeking self-promotion can engage others in the charitable endeavour, thereby offering a prosocial rationale for their self-promotion and mitigating doubts about their motives (O'Connor et al., 2020).

### **1.1.3 The Cognitive Mechanisms Supporting Reputation-Based Cooperation**

The previous paragraphs have established that for reputation-based cooperation, people must both be able to evaluate the character and motives of others, as well as strategically manage their own reputation. The latter skill likely involves more complex mechanisms. While some of these mechanisms have been touched upon in previous paragraphs, the following will more explicitly consider various psychological adaptations underlying reputation-based cooperation (for a comprehensive examination of the topic see Manrique et al., 2021).

Manrique et al. (2021) argue that an extensive working memory is a crucial social-cognitive mechanism, which is linked to three further socio-cognitive abilities: understanding others' mental states, adhering to and enforcing social norms, as well as delaying gratification. Working memory is for instance likely to play a role in assessing the reputations of others, such as when monitoring cooperative actions (Milinski & Wedekind, 1998). Tracking one's own appearance to others necessitates the ability to consider multiple perspectives simultaneously, which places significant demands on working memory (Manrique & Walker, 2017). Additionally, effectively managing one's own reputation may involve mental time travel, enabling individuals to envision future scenarios, which likely further engages working memory (Dere et al., 2019).

Theory of Mind is a complex concept that entails the capacity to attribute mental states to oneself and others, encompassing various cognitive abilities. While reputation management itself may not necessitate Theory of Mind, possessing Theory of Mind introduces significant strategic complexity by enabling individuals to tailor their reputation management based on the knowledge of observers. Individuals should possess psychological mechanisms not only for monitoring the likelihood of being observed but also for considering various factors such as the characteristics of the potential audience, the perceived value of one's reputation to that audience, the anticipated impact of specific actions on how the audience perceives oneself, and even strategies for avoiding undesirable behaviours from being observed (Barclay, 2013). Moreover, reputations often involve implicit comparisons with others, such as being perceived as more

cooperative than peers. Consequently, people should also possess capacities to track their reputation relative to others and adjusting their behaviour accordingly (Barclay, 2013).

Intention attribution and perspective-taking are two distinct abilities that are likely to play a role in both reputation management and evaluation. As established in the previous section, it is important to be able to correctly infer the intentions behind others' actions as well as grasp how one's own actions are perceived by others. Individuals do not only fail to cooperate for selfish reasons, but may also do so because they for instance lack the resources / ability or were not aware another required assistance. If observers can correctly interpret uncooperative actions as an error or short-term lack of ability, they can maintain a cooperative relationship with genuinely helpful individuals. Consequently, the reputation system becomes less susceptible to errors that could threaten sustained cooperation. While some intentions are straightforward or openly communicated, individuals occasionally intentionally conceal their intentions to deceive others. The ability to discern hidden intentions may be confined to humans, as it may necessitate a full-fledged Theory of Mind and a robust working memory capable of simultaneously juggling multiple perspectives (Manrique & Walker, 2017). Epistemic vigilance, described earlier as a set of cognitive abilities to distinguish trustworthy information from misinformation, encompasses the capacity to scrutinise intentions that seem contradictory to observed behaviours. While non-human apes and young children can attribute intentions that align with an observed behaviour (Tomasello et al., 2005), the evolution of scepticism entails a more complex process. We are not only able to realise that an individual's intentions contradict their actions but also recognise that the individual desires us to incorrectly infer their motives (such as when someone donates to gain reputation benefits rather than to be genuinely helpful). This understanding entails employing advanced meta-representational reasoning (Sperber et al., 2010), which could explain why it typically develops later in childhood (Heyman et al., 2007; Heyman et al., 2014).

Perspective-taking is necessary to both manage one's reputation and assess the reputations of others. A key requisite for reputation management is the skill to anticipate how one's behaviours will be perceived by others. Thus, adopting others' viewpoints can significantly enhance an individual's ability to manage their reputation. Similarly, perspective-taking enhances the ability to identify non-cooperators: individuals may deceitfully portray themselves as cooperative, and it demands cognitive effort for observers to discern between true and false cooperators. Besides the skills mentioned in previous paragraphs, it also requires the capacity to simultaneously interpret both overt and concealed intentions in others' behaviours.



In order to establish a favourable reputation and accurately assess the reputations of others, it is imperative to understand what the societal norms are, and whether behaviours adhere to them. Without a clear understanding of what constitutes acceptable behaviour, it becomes challenging to discern suitable partners based on their adherence to these norms. As explained in previous paragraphs on IR models, cooperation necessitates strategies that evaluate actions based on the perceived reputation of the recipient, which requires the establishment of social norms dictating when assistance should (or should not) be provided (Ohtsuki & Iwasa, 2007; Kandori, 1992). While RBPC can theoretically operate without norms (e.g., when observers only assess an actor's capability to assist), norms can nonetheless influence these decisions. The same act of assistance may be viewed as generous if it aligns with prevailing norms of limited help, or as stingy if it contradicts norms advocating for greater assistance (Barclay, 2013). Comparing potential partners to established norms can aid in decision-making (McNamara et al., 2008), as can comparing one's own behaviour to these norms and adjusting cooperation levels accordingly (Barclay, 2013, 2016). In line with this, children's ability to manage their reputation appears to develop after they grasp the concept of norms as collectively agreed-upon standards for cooperation (Engelmann et al., 2012; Kelsey et al., 2018).

In addition, engaging in reputation management, particularly when it involves costly cooperation, requires the capacity to resist immediate gains, such as cheating, and to prioritise larger future benefits, recognising that rewards may originate from future partners, rendering them delayed and more uncertain. Indeed, research suggests that children's capacity to delay gratification correlates positively with their inclination to share, suggesting that the ability to delay gratification may serve as a foundational aspect for children's tendencies towards sharing and cooperation (Sebastián-Enesco & Warneken, 2015). Moreover, in adults, directing attention towards the future has been shown to foster greater generosity (Sjåstad, 2019) and to encourage them to willingly bear costs to avert the spread of detrimental reputational information (Vonasch et al., 2018).

#### **1.1.4 A Developmental Approach to Impression Formation and Management**

While this thesis focuses on adult participants, understanding the developmental trajectory of impression formation and management can further highlight the cognitive mechanisms

underlying these processes. This section will describe how reputation-related abilities emerge in childhood, setting the stage for their role in adulthood.

In early childhood, children acquire the capacity to both cultivate their own reputation and assess the reputations of others (Banerjee et al., 2020; Engelmann & Rapp, 2018; Silver & Shaw, 2018). Hereby, the ability to perceive traits in others ostensibly emerges before children learn the skill of managing their own reputation. Before reaching the age of three, children demonstrate an ability to differentiate between friendly and unfriendly partners solely through observation, and show a preference for friendly individuals (Herrmann et al., 2013). At three years old, children anticipate individuals sharing information that portrays them positively, such as showcasing achievements while concealing failures (Hicks et al., 2015). At four years old, they understand that individuals may lie for personal gain, and by age six, they can discern between lies driven by self-interest and those aimed at benefiting others (Banerjee et al., 2020; Mills & Keil, 2005; Shaw & Olson, 2015). From the age of five, children also selectively choose to cooperate with individuals who possess task-relevant qualities whilst avoiding them as competitors (Grueneisen et al., 2023), and engage in gossip providing their assessments of others to aid third parties in deciding whom to interact with (Engelmann et al., 2016).

The emergence of strategic reputation management appears to occur later in development. While young children under 2 years old demonstrate prosocial behaviour, it seems driven more by an intrinsic motivation to meet others' needs rather than strategic reputation management (Dunfield et al., 2011; Vaish et al., 2009; Warneken & Tomasello, 2007). Around the age of 5, children begin to adopt self-promotional strategies (reviewed in Engelmann & Rapp, 2018) and actively seek to maintain a positive reputation (Fu et al., 2016). Five-year olds for instance refrain from behaviours like stealing when they know they are being observed or make more generous donations when their actions will be known by others (Grueneisen & Tomasello, 2017; Leimgruber et al., 2012; McAuliffe et al., 2020). By approximately age 8, children begin explicitly referring to self-presentational concerns when explaining their behaviour in front of an audience (Banerjee et al., 2012). For instance, in a study by Aloise-Young (1993), eight-year-olds demonstrated appropriate verbal self-presentational strategies to increase their chances of being chosen as a game partner. Moreover, children around this age exhibit flexibility in tailoring their reputational strategies—whether through self-promotion, ingratiation, modesty, or disclaimers—to cultivate a good reputation (Watling & Banerjee, 2012). Correspondingly, they become sensitive to self-presentation motives when evaluating others' prosocial actions (Amemiya et al., 2021; Banerjee, 2002; Heyman et al., 2014). This

ability continues to develop over time, with children aged nine to ten understanding that self-promotional statements are more morally acceptable when they can, for instance, be interpreted as implicit offers of assistance (Amemiya et al., 2021).

Furthermore, the tendency to engage in strategic prosocial behaviour in order to elicit reciprocal responses from others begins to emerge around the ages of four to five, developing in complexity over subsequent years (reviewed in Grueneisen & Warneken, 2022). For instance, in a study where children aged three to seven interacted with a 'game owner' who had the option to select either the child or a competitor as a play partner for an appealing game (with a control condition where the game owner played alone), the child and competitor were first given the opportunity to share either a more desirable or a less desirable sticker with the game owner (Warneken et al., 2019). While three-year-olds exhibited indiscriminate sharing, children aged 5 and older were more inclined to share the attractive sticker in the experimental condition compared to the control. Eight-year-olds also demonstrated competitive altruism, wherein they strove to outdo each other in sharing resources when observed by third parties who might later include them in an enjoyable activity (Herrmann et al., 2019). Therefore, it appears that the refinement of social skills necessary for individuals to form impressions of others, understand how their actions shape their reputations and strategically cultivate a positive image, unfolds gradually throughout childhood.

## **1.2 Punishment and Reputation**

Another mechanisms that has been proposed as key in supporting the evolution of cooperation among non-relatives is punishment, an act that lowers the immediate payoff of actor and target (Arseneau-Robar et al., 2016; Fehr & Gächter, 2000, 2002; Raihani & Bshary, 2019; Raihani et al., 2010, 2012). Punishment has been observed across diverse societies (Henrich et al., 2010; Henrich et al., 2006), emerging early in development (Bašić et al., 2020; Kanakogi et al., 2022; Koymen et al., 2014; Marshall & McAuliffe, 2022; Marshall et al., 2021; McAuliffe et al., 2015; Rossano et al., 2011; Schmidt et al., 2012; Yudkin et al., 2020), thereby underscoring punitive actions as a fundamental aspect of human social interaction. Human punishment has often been studied within the framework of stylised laboratory games (Balliet & Van Lange, 2013; Balliet et al., 2011; Chaudhuri, 2011; Dreber et al., 2008; Fehr & Gächter, 2000, 2002; Henrich et al., 2012; Nikiforakis & Normann, 2008; Raihani & Bshary, 2019; Raihani et al., 2012). In these games, participants are typically endowed with money that they can invest in

collective endeavours. Participants may also opt to defect by retaining the money for personal gain or by exploiting the contributions of their group members. Punishment mechanisms are integrated into these games by offering participants the choice to pay a nominal 'fee' to impose a (usually) more substantial 'fine' on a target. Punishment can promote cooperation where individuals would otherwise be tempted to cheat (Clutton-Brock & Barker, 1995; Raihani et al., 2012; Yamagishi, 1986). However, this alone cannot explain the evolution of punishment, especially since punishment – like cooperation – imposes immediate costs on the punisher. Punishing a free rider can be considered a prosocial behaviour because the entire group can benefit if the free rider subsequently cooperates more (Yamagishi, 1986). Furthermore, those who choose not to punish benefit more than those who do, as they reap the rewards of punishment without incurring the costs (Barclay, 2006). Punishment is sometimes referred to as a 'second-order public good', susceptible to exploitation by second-order free-riders who cooperate but refrain from punishing defectors (Boyd et al. 2003; Boyd & Richerson 1992). Especially when individuals engage in repeated interactions with the same partner (Dreber et al., 2008; Rand et al., 2009; Wu et al., 2009) or when individuals can punish others for harming a third party (Ohtsuki et al., 2009; Ule et al., 2009), those who choose to enforce punishment typically end up with lower payoffs compared to those who do not.

While punishment can lead to more cooperative behaviour from the target in future interactions, this alone is unlikely to be the primary driving force behind the natural selection of punishment. This assertion is supported by the scarcity of such outcomes observed in experimental settings (as reviewed in Raihani & Bshary, 2019). Furthermore, both laboratory studies (e.g., Fehr & Fischbacher, 2004) and field experiments (e.g., Balafoutas & Nikiforakis, 2012; Matthew & Boyd, 2011) have demonstrated that individuals sometimes engage in punishment even when they act as third-parties, having not been directly victimised by the transgression and not anticipating future interactions with either the victim or the wrongdoer. Consequently, punishment may have evolved for purposes other than merely promoting cooperation (Dreber et al., 2008; Ohtsuki et al., 2009; Raihani & Bshary, 2019; Rand et al., 2009; Ule et al., 2009; Wu et al., 2009). While punishers can be motivated by a desire for retribution (Carlsmith et al., 2002), as well as concerns regarding relative payoffs, such as avoiding having less than others (i.e., disadvantageous inequity aversion; Raihani & Bshary, 2019; Raihani & McAuliffe, 2012) and desiring equal outcomes for everyone (i.e., general egalitarian preferences, Dawes et al., 2007), individuals also punish for competitive reasons (e.g. to enhance their position in comparison to others; Raihani & Bshary, 2019). Punishment is recognised as a versatile behavioural instrument capable of fulfilling diverse functions that are not necessarily mutually

exclusive (Raihani & Bshary, 2019). Crucially, a prominently proposed avenue for punishers to regain their investment lies in reputation: the influence of reputation can significantly bolster the chances of punishers participating in advantageous social interactions in subsequent contexts (Barclay, 2006; Batistoni et al., 2022; dos Santos & Wedekind, 2015; dos Santos et al., 2011, 2013; Jordan et al., 2016a; Raihani & Bshary, 2015b,c; Redhead et al., 2021). In this thesis, it is important to acknowledge that the principles (e.g. IC, RBPC) that apply to reputations in one domain – such as cooperation - often extend to reputations in other domains, such as punishment (Barclay, 2015).

Establishing a reputation as a punisher can offer advantages via two main routes (Raihani & Bshary, 2015b): (i) it can signal formidability, deterring interaction partners or observers from defecting in future (e.g. dos Santos et al., 2013; Hilbe & Traulsen, 2012); or (ii) it can signal a willingness to cooperate, thereby granting punishers increased opportunities for cooperative engagements with new social partners (e.g. Barclay, 2006; Dhaliwal et al., 2020; Jordan et al., 2016; Nelissen, 2008; Raihani & Bshary, 2015b). Punitive actions can be seen as signals that enable the punisher to communicate an otherwise hidden intention to cooperate (Jordan & Rand, 2017; Jordan et al., 2016a; Przepiorka & Liebe, 2016). When the display of this signal (punishment) is linked with the underlying trait (cooperativeness), observers can respond accordingly to the informative nature of the signal. They may be more inclined to reward punishers (Raihani & Bshary, 2015c), trust them (Barclay, 2006; Jordan et al., 2016a), or prefer them as partners over non-punishers (Batistoni et al., 2022; Raihani & Bshary, 2015b). However, the empirical evidence surrounding how punishment affects reputation lacks clear consensus. Although various studies align with the theoretical models suggesting that punishers benefit from a punitive reputation because it deters future partners from defecting (e.g., Fehr & Fischbacher, 2003; dos Santos et al., 2013), other research shows more contradictions. While punishers are trusted in certain contexts (Barclay, 2006; Jordan et al., 2016a; Nelissen, 2008), they are not always liked or rewarded (Balafoutas et al., 2014; dos Santos et al., 2013; Horita, 2010; Kiyonari & Barclay, 2008; Ozono & Watabe, 2012). Additionally, punishers sometimes display (Kurzban et al., 2007; Piazza & Bering, 2008) and sometimes conceal (Rockenbach & Milinski, 2011) their punitive actions: advertising one's punitive actions suggests that it leads to reputation advantages, while hiding them implies the opposite. The inference therefore is that the information conveyed by punishment and its resulting reputation effects are likely to depend on the specific context (Raihani & Bshary, 2015b).

In order to grasp the reputational ramifications of punishment, it becomes imperative to first ascertain the type of information observers can deduce regarding the underlying motive behind the decision to punish. However, inferring the motives behind punitive actions is challenging, as punishment inherently involves causing harm. Punishment can originate from vindictive and competitive intentions or from a sense of fairness and a commitment to reinforcing societal norms (Brañas-Garza et al., 2014; Dhaliwal et al., 2020; Raihani & Bshary, 2019), with some scenarios allowing clearer evaluations than others. Consequently, the extent to which punishment is reflective of an individual's cooperative inclinations varies (Brañas-Garza et al., 2014; Nelissen, 2008; Raihani & Bshary, 2015b). Approval of punishers may be confined to specific circumstances where competitive or self-serving motivations can be discounted (Barclay, 2006; Jordan et al., 2016; Kiyonari & Barclay, 2008; Raihani & Bshary, 2015b, 2015c). This is proposed to occur when punishment targets a cheat rather than a cooperator, thus justifying the punitive action; when the punisher themselves was not the victim of the transgression; and when alternative methods of directly signalling cooperative intentions are absent (Raihani & Bshary, 2015b). Situations in which the punisher is not the victim of a transgression and does not directly benefit from changes in the transgressor's behaviour is generally observed in cases of third-party punishment (TPP). TPP refers to scenarios in which the punisher was not the victim, while second-party punishment (SPP) refers to situations in which the punisher was (one of) the victim(s) of the transgressor. TPP has commonly been explored within the framework of one-shot games, where the punisher has the option to penalise a defector who has harmed someone else (e.g., Kurzban et al., 2007; Nelissen, 2008; Piazza & Bering, 2008). As the third party is not directly affected by the cheat and will not engage with them in subsequent interactions, onlookers may interpret the punishment as an expression of concern for the victim's well-being. Conversely, when the punisher themselves has been victimised by the cheat, bystanders may perceive the punitive action as stemming from the punisher's response to loss or injustice resulting from their interaction with the cheat (e.g., Fehr & Gächter, 2002; Raihani & McAuliffe, 2012). In these instances, punishment is less likely to convey genuine concern for others. In fact, antisocial and SPP are positively associated with competitive motives rather than cooperative ones (Carpenter, 2003; Espin et al., 2012; Peysakhovic et al., 2014; Yamagishi et al., 2012). Conversely, SPP has also been construed as driven by a desire to evade a subordinate status (Yamagishi et al., 2012). Analogous to how cultivating a punitive reputation can confer advantages (Barclay, 2006; dos Santos et al., 2011, 2013; dos Santos & Wedekind, 2015; Raihani & Bshary, 2015b, 2015c), acquiring a reputation

for being submissive may prove detrimental (Cohen et al., 1996; Crombag et al., 2003; Osgood, 2017).

Similarly, punishment is a more ambiguous signal of cooperative intent compared to helping, as acts of assistance are less likely to originate from harmful or competitive motives (but see Berman & Silver, 2022; Raihani & Power, 2021). Individuals who opt for third-party punishment over third-party compensation often exhibit higher levels of antisocial personality traits, including Machiavellianism, narcissism, and psychopathy (Dhaliwal et al., 2021). Thus, when individuals are presented with the choice between compensating a victim or punishing a transgressor, punishment is less apt to convey cooperative intent (Batistoni et al., 2022; Dhaliwal et al., 2021; Heffner & FeldmanHall, 2019; Jordan et al., 2016a; Raihani & Bshary, 2015c). Those who reward prosocial behaviour or compensate victims tend to receive more favourable evaluations from observers and are more likely to be selected as potential interaction partners compared to punishers (Dhaliwal et al., 2021; Lee & Warneken, 2020; Ozono & Watabe, 2012). Actors appear cognisant of the reputational consequences of these actions and generally choose to reward prosocial behaviour or compensate victims over punishing transgressors when given the choice (Heffner & FeldmanHall, 2019; Jordan et al., 2016b; Molenmaker et al., 2016). Inaction produces opposite reputation effects: recent findings indicate that non-helpers are perceived as less trustworthy, whereas there is no discernible effect for non-punishers (Batistoni et al., 2022). Hence, while punishers are less likely to accrue a positive reputation than helpers, failing to provide assistance carries greater reputational repercussions than failing to administer punishment (Batistoni et al., 2022). Additionally, it is noteworthy that in certain circumstances, such as when observers are selecting delegates to deter potential exploitation, punishers may be preferred over helpers (Li et al., 2023).

Other factors that punishment signalling is sensitive to are deservedness and severity (Salcedo & Jimenez-Leal, 2024), as well as fairness (Bone et al., 2016; Deutchman et al., 2021; Raihani & McAuliffe, 2012), blameworthiness (Buckholtz et al., 2008, 2015), and outcomes (Cushman et al., 2009; Martin & Cushman, 2016; Mazzocco et al., 2004). The ‘efficiency’ of punishment can also influence observer perceptions. Empirical studies commonly employ a fee-to-fine ratio larger than one (e.g., Barclay, 2006; Chavez & Bicchieri, 2013; dos Santos et al., 2013; Dreber et al., 2008; Fehr et al., 2012; Fehr & Gächter, 2002; Herrmann et al., 2008; Horita, 2010; Kiyonari & Barclay, 2008; Kurzban et al., 2007; Nelissen, 2008; Nikiforakis & Engelmann, 2011; Ozono & Watabe M, 2012; Piazza & Bering, 2008; Rockenback & Milinski, 2011), meaning that the cost of punishment is lower than the cost it inflicts on the target. As such,

punishment improves the punisher's payoff's relative to those of the target, and punishers in economic games may face observers' doubt concerning the moral value of the act (Fehr & Rockenbach, 2003; Xiao, 2013). Even TPP can be perceived as stemming from competitive motives (e.g. Pedersen et al., 2013). In fact, third-party punishers who bear significant costs to punish transgressors are preferred over those who bear lower costs (Nelissen, 2008), aligning with the notion that observers consider the relative cost of punishment when evaluating the punisher. Further, a meta-analysis revealed that punishment tends to foster cooperation more effectively when the actor incurs a cost, as opposed to when punishing is cost-free (Balliet et al. 2011).

Individual differences among punishers can also shape perceptions of their actions. Those with higher levels of power or social status, who are less likely to face retaliation, are expected to engage in punishment (Gordon & Lea, 2016), and indeed demonstrate a greater willingness to penalise transgressors (Molho et al., 2020; Redhead et al., 2021). Experimental findings further suggest that individuals strategically punish to signal formidability, instilling enough fear to dissuade others from mistreating them in future interactions (Krasnow et al., 2016), and that punishers are indeed perceived as more dominant (Gordon et al., 2014; Gordon & Lea, 2016). Another contextual factor that can impact a punisher's reputation is the identity of the individual being punished. A recent study, for example, revealed that TPP enhanced perceived trustworthiness when directed at an out-group member, as it signals integrity (Sun et al., 2023). Conversely, bystanders perceived punishers as less trustworthy when they administered punishment to an in-group member, stemming from a reduction in benevolence consciousness. The original violation further influences punishment. Across 15 diverse populations, a proportional adjustment of punishment to the severity of offenses in economic games was found, with a greater number of individuals inclined to punish as behaviours became increasingly unfair (Henrich et al., 2006). This observation aligns with similar findings reported in Western samples (Fehr & Fischbacher, 2004). In a similar vein, a vignette study conducted in small-scale and Western societies consistently indicated that punishment is perceived as more appropriate when offences are intentional rather than unintentional (Barrett et al., 2016).

In the commonly employed economic game setup, in which punishment is administered via monetary fines without clarification regarding the rationale behind the punishment or consensus on the actions meriting punishment, there is considerable room for variation and misinterpretation when targets attempt to discern a punisher's motivations. Moreover, punitive behaviours observed in laboratory settings frequently diverge from those witnessed in real-life



(Balafoutas et al., 2014; Pedersen et al., 2020). In the real world, people are for instance less willing to punish when opportunities to do so arise (Balafoutas et al., 2014; Baumard, 2010; Guala, 2012; Pedersen et al., 2018; but see Mathew & Boyd 2011). This disparity arises because people actually possess a range of intervention options, encompassing actions like direct physical or verbal confrontation, withholding help rather than actively harming transgressors (Balafoutas et al. 2014), alongside indirect strategies such as social avoidance (Boehm 2001; Lee, 1979) and gossip (Molho et al., 2020). Gossip refers to conversations that convey social information about individuals who are not present. It serves as a means to exchange or collect information about others in the absence of direct observation, and individuals are aware of its potential impact on their reputation, exhibiting more prosocial behaviours when reputation is at stake (Feinberg et al. 2012, 2014; Jolly & Chang 2021; Wu et al. 2016a, 2016b). Gossip is pervasive in real-world contexts and can be used to inflict reputation costs on transgressors (Beersma & Van Kleef 2012; Dunbar 2004). Furthermore, it shares a similar payoff structure to punishment of economic games: it incurs costs both for those perceived as engaging in gossip (Adams et al., 2012) and for those who become the subject of gossip (Feinberg et al., 2012). Tapping into this, Chen and Xu (2020) argue that contradictory findings surrounding the reputation of punishers are a result of studies treating reputation as a uni-dimensional variable that is negative or positive, failing to appropriately consider punishment motives, and over-focusing on financial punishment whilst neglecting other forms of punishment.

Thus, punishment as observed “in the real world” often diverges from the structure of commonly used economic game. In settings that are more ecologically valid, a multitude of factors—relational, situational, and emotional—can influence the dynamics of punishment (Molho et al., 2020). Ethnographic records, for instance, illustrate that punishment frequently manifests as criticism or mocking (Baumard 2010; Boehm 2011; Wiessner 2005). For example, among the Ju/'Hoansi bushmen, rather than resorting to physical aggression, conflict mediation and expressions of displeasure towards others involve a range of reputational put-downs, such as mockery, ridicule, and direct criticism (Wiessner, 2005). Overall, the motivation ambiguity and context-dependency inherent in punishment highlight the necessity of moving beyond the fundamental and numerous insights provided from the decontextualized settings often utilised in economic games to more ecologically valid environments. Doing so is crucial for gaining a comprehensive understanding of the factors that influence a punisher's reputation.

In general, a lack of ecological validity occurs when a research setting does not accurately represent the natural environment of the individuals being studied (Brunswik, 1956). As a

result, it becomes challenging to draw conclusions about how people would behave in their real-world context based on their behaviour observed in the research setting (Brunswik, 1956). More specifically, ecological validity is a particular form of external validity. It refers to the degree to which research findings can be applied to natural, everyday environments (Krolak-Schwerdt et al. 2018). Generally, there is a trade-off between the level of experimental control and ecological validity, and the significance of this balance has been a topic of ongoing debate (Albright & Malloy, 2000). Studies with greater control tend to have lower ecological validity, making it harder to generalise their results to real-world settings (Krolak-Schwerdt et al. 2018).

### **1.3 Relevant Methodological Frameworks**

The following section will briefly describe relevant methodological frameworks, which vary in their levels of ecological validity and experimental control.

The development of economic games, which have become a well-established tool for studying social behaviour (Freedman & Flanagan, 2017; van Dijk & De Dreu, 2021), represents a major advancement in the study of human prosociality. Originating in behavioural economics, these paradigms have become the standard for assessing prosocial behaviours in controlled, experimental settings (Baumard et al., 2013). Economic games are designed to establish strong causal links, enabling researchers to explore the relationship between experimental manipulations and participant behaviour. To ensure this, the games are crafted to be highly internally valid, meaning they are standardised and repeatable (Bardsley et al., 2010; Guala & Mittone, 2010; Loewenstein, 1999). They tend to be theory-driven and simplify real-life situations and challenges to their essential features, in hopes of offering a coherent and substantive model of various real-world encounters (Murnighan & Wang, 2016). These games aim to reveal people's strategies that reflect their choices in analogous social interactions (Murnighan & Wang, 2016). In many economic games, participants receive sums of money equivalent to a few minutes' work up to a day's wages (Larney et al., 2019). Usually, participants (players) must then decide how much to keep for themselves and how much to allocate to others.

One advantage of economic games lies in their ability to incorporate "real" outcomes, generally in the form of monetary bonus payments, thereby making decisions genuinely consequential. Unlike hypothetical scenarios or self-reported questionnaires, where reporting behaviour may not accurately reflect one's true preferences, in economic games, deviating from genuine preferences can come at a cost. Failing to carry out such studies 'properly' is more likely to

result in forfeiting preferred outcomes. Consequently, participants are generally expected to be more motivated and attentive (Baron, 2001; Camerer & Hogarth, 1999). Moreover, responses in economic games are thought to be less susceptible to the influence of social desirability bias—the tendency to present oneself in a socially desirable manner rather than truthfully. While individuals can simply claim to be prosocial in a questionnaire, players in incentivised games must tangibly back their claims with actions. To appear more prosocial than they genuinely are, players must sacrifice desired outcomes. Therefore, the presence of consequences in economic games transforms observations into instances of genuine behaviour (Baumeister et al., 2007; Klein & Hilbig, 2019; Murnighan & Wang, 2016; Pruitt & Kimmel, 1977). Similarly, it has been argued that the results generated by economic games serve as the direct variable of interest, representing the construct to be measured (preference) without necessitating extensive additional assumptions (Thielmann et al., 2021). Economic games have therefore been likened with the behaviourist tradition: akin to classical conditioning experiments, the strength of the conditioned response directly mirrors the construct of interest (i.e., the association between stimuli) (Thielmann et al., 2021). While the causal mechanisms driving these associations are distinct, the paradigm supposedly maintains an element of objectivity—observed responses themselves serve as the construct of interest.

In Chapter 6 of the current thesis, I use an adaptation of a commonly used economic game: the “Trust Game” (Berg et al., 1995). This game is a standard laboratory experiment for measuring trust. The Trust Game is a sequential game with two players in asymmetric roles, referred to as the ‘Sender’ (also called trustor or investor) and the ‘Receiver’ (also called trustee). Both players (or only the Sender) receive a monetary endowment, which the Sender can choose to transfer in whole or in part to the Receiver. The Sender retains any portion of the endowment they do not transfer. Typically, the experimenter triples the transferred amount before it reaches the Receiver. Subsequently, the Receiver can decide to return some, all, or none of the received transfer. The Sender's actions are viewed as an indication of trust, while the Receiver's actions are seen as a demonstration of trustworthiness (Camerer, 2003).

Contrary to their strong internal validity, economic games have faced criticism for their limited external validity, or applicability to contexts beyond the experimental setting (Baumard & Sperber, 2010; Galizzi & Navarro-Martinez, 2019; Hagen & Hammerstein, 2006; Loewenstein, 1999; Wiessner, 2009). Some argue that the principles and norms underlying classical economic games may not hold true outside of large-scale market economies (Baumard & Sperber, 2010; Wiessner, 2009) or may not address theoretical inquiries beyond bargaining (Hagen &

Hammerstein, 2006; Loewenstein, 1999). Overall, there is a growing consensus that researchers should not rely solely on the conventional set of economic games merely because they are considered "standardised" (Guala & Mittone, 2010) or "paradigmatic" (Guala, 2008; Guala & Mittone, 2010) (but refer to Pisor et al., 2020 for arguments against abandoning economic games as a methodological tool based on these critiques, along with examples of how researchers can strategically adjust the framing or design of economic games to infer real-world preferences). Furthermore, to minimise bias in deciders' decisions, many classical economic games provide participants with simplified task descriptions, avoiding real-world analogies (Friedman & Cassar, 2004). Outcomes derived from economic games are highly sensitive to how the experiment is framed (Kahneman, 2003; Lightner et al., 2017). Results from games framed with real-world analogies can significantly diverge from those without such framing (Cronk, 2007; Lessorogol, 2005; Lightner et al., 2017). Importantly, when instructions are minimalist, it becomes uncertain which real-world analogies or personal experiences participants rely on to guide their decision-making (Baumard & Sperber, 2010; Ensminger, 2000; Gervais, 2017; Hagen & Hammerstein, 2006; Pisor & Fessler, 2012), which complicates interpretations of the data.

While economic games, by their decontextualised nature, can for instance unveil judgments and preferences regarding interactions with punishers and helpers, it is important to note that these findings may not apply to real-world contexts. Batistoni et al. (2022) for instance suggested that although participants in their study tended to view help as more appropriate than punishment, this preference might not hold true in other situations where punishment is perceived as more suitable than compensating victims (Reb et al., 2006). Additionally, there are instances where punishers may be endorsed and preferred over helpers in interactions, particularly in evaluations related to the attributes of a strong and capable leader or in hypothetical leadership selections (Dhaliwal et al., 2021). Therefore, it is crucial to reiterate what was established in preceding sections: context matters. For instance, experimental research often assumes perfect monitoring, where people can observe each other's behaviour and appropriately reward or punish (e.g., Fehr & Gächter, 2002; Ozono et al., 2020). However, real-life social interactions are often marred by 'noise'. Studies investigating how leaders' reputations are influenced by noise-induced errors in punishment or reward found that mistaken punishment harms a leader's reputation, whereas mistaken reward does not (de Kwaadsteniet et al., 2019). This discrepancy may stem from the inherently harmful nature of punishment, leading to more negative judgments when applied erroneously compared to rewards. Alternatively, uncertainty in judging a punisher's motives under noise may prevent reputation benefits unless punishment is

perceived as justified and motivated by the right intentions. To gain deeper insights into the reputation consequences of cooperation and punishment in real-world settings, it is essential for research to move beyond economic games and models and examine reputation under noisy, intricate conditions - an area that has been comparatively neglected in psychology.

In many experimental games, the decision to cooperate or defect is often a binary choice, where players learn about each other's choices afterward. However, interactions with strangers in our evolutionary past were unlikely to involve the exchange of 'closed boxes' where people only discovered each other's contributions after their own choice was irrevocably made. Actual situations likely entailed simultaneous or phased exchanges, allowing individuals to monitor each other's behaviour, make real-time adjustments, and, importantly, end exploitative exchanges right away (Connor, 1992; Hart & Hart, 1992). When interactions involve multiple decisions, stable cooperation can emerge even between strangers without a long-term perspective (Raihani & Bshary, 2015a). While focusing on a set of simplified assumptions and standardised games renders modelling and economic game outcomes more manageable, they also deviate significantly from the intricate process of constructing, evaluating, and leveraging reputations in real-world scenarios. A study examining the link between cooperative behaviours and prosocial reputations in an Afro-Caribbean village for instance found that simple models of indirect reciprocity do not capture the actual process of reputational formation (Macfarlan et al., 2013). Instead of following numbers of cooperative acts, prosocial reputations were correlated with the number of individuals assisted, and while reputations followed cooperative behaviours, they were anchored across time (Macfarlan et al., 2013).

In summary, while controlled experiments are valuable for testing theories with internal validity, theories should not focus exclusively on behaviour observed in laboratory settings. Instead, they should aim to explain behaviours as they naturally occur in daily life across diverse situations and relationships. Nevertheless, most psychology research on human prosocial and punitive behaviours has been conducted in artificial, abstract laboratory environments, such as economic games (Balliet et al., 2017; Spadaro et al., 2022; Thielmann et al., 2021). Empirical research on any natural phenomenon must document it within its natural context to determine what the theory needs to explain and to guide the development of appropriate theories (Guala, 2012; Rozin, 2001, 2006). Balliet et al. (2022) further advise that research can be enhanced by closely examining the phenomena that theories aim to explain.

Two effective methods for achieving this are experience sampling and diary methods (Hofmann et al., 2014; Mehl & Conner, 2011). The experience sampling methodology (ESM) is a data

collection method in which participants respond to repeated assessments over time while in their natural settings. This approach allows researchers to study social dynamics over extended periods, capturing behaviours that cannot be observed in most laboratory settings. With experience sampling, researchers can contact participants multiple times throughout the day and have participants report on social behaviours they have recently experienced across diverse real-life situations and relationships, which can provide a test of theory and/or identify variation in phenomena that theory should explain. Similarly, diary methods involve asking participants to report on their daily experiences at the end of each day (Bolger et al., 2003). Both methods reduce recall bias and the use of global heuristics by collecting data close to the time events occur, thereby providing more accurate self-reports of actions and feelings (Schwarz, 2012; Scollon et al., 2009). Most relevantly, a key strength of ESM is that it takes psychological research out of the laboratory and into real-life contexts, enhancing ecological validity (Scollon et al., 2003). It also helps researchers gain a deeper understanding of the factors influencing behaviours and enables the examination of within-person processes (Scollon et al., 2003). Additionally, ESM methods let researchers compare participants' behaviours in laboratory tasks with their reported behaviours in daily life.

A recent study by Weiss et al. (2021), for instance, compared trust in Trust Games (Berg et al., 1995) to trust in daily life. This study found that behaviour in the Trust Game predicted trust in strangers but not in close relationship partners. Similarly, another study using ESM found that direct punishment, commonly observed in laboratory settings, is rare in real-life interactions between strangers (Molho et al., 2020). Additionally, high-conflict situations among strangers constitute only a small fraction of what people actually experience in daily life (Columbus et al., 2021). These findings highlight the limitations of laboratory studies, which often focus on interactions between strangers, and underscore the importance of using methods like ESM.

Chapters 2 and 3 of this thesis draw inspiration from ESM. In Chapter 2, I investigate reputation management in everyday life using a daily diary phone app. Chapter 3 comprises a weekly survey study, incorporating repeated measures over six weeks, to assess the reputation of punishers in daily life. Both studies delve into the complexities of real-life situations, considering multiple situational factors and relationship contexts to effectively study participants in their natural settings.

To examine whether actors are aware of how their cooperative, uncooperative, and punitive actions affect their reputation, and to determine whether observers can accurately attribute motivations to actors, studies in Chapters 4 and 5 employed a componential approach to

modelling judgment accuracy, drawing from existing literature on social judgment accuracy (Biesanz, 2010; Funder, 1995; Kenny, 2004; West & Kenny, 2011). Crucially, rather than relying on hypotheticals, data collection involved soliciting written descriptions of recent cooperative, uncooperative, and punitive behaviours along with accompanying self- and meta-perceptions. Subsequently, these written accounts of recent behaviours were provided to a separate set of judges. These judges then made actual judgments on the same items. This design allowed for a direct assessment of actors' meta-accuracy, as well as an examination of the extent to which observers can predict the self-reported motives and traits of actors in the context of real-world cooperative, uncooperative and punitive behaviours.

As previously noted, experimental designs have often prioritised internal validity at the expense of external validity and generalisability (Argyris, 1975; Scandura & Williams, 2000). Other methods, such as ESM, sacrifice causality. It has been argued that the experimental vignette methodology (EVM) offers a solution to this dilemma (Aguinis & Bradley, 2014). EVM involves presenting participants with carefully constructed and realistic hypothetical scenarios to assess dependent variables such as intentions, attitudes, and behaviours (Alexander & Becker, 1978). This approach enhances experimental realism while also allowing researchers to manipulate and control independent variables, thus simultaneously bolstering both internal and external validity (Atzmüller & Steiner, 2010; Hox et al., 1991; Krolak-Schwerdt et al. 2018) (for articles addressing important aspects of EVM, refer to Aiman-Smith et al., 2002; Graham & Cable, 2001; Hughes & Huby, 2002; Ludwick & Zeller, 2001; Shepherd & Zacharakis, 1999; Wason et al., 2002; Wilks, 2004). Specifically, vignettes can offer a standardised stimulus to all participants, thereby enhancing internal validity, measurement reliability, and facilitating replication (Wason et al., 2002). Further, studies across various domains have successfully validated vignettes against real-world behaviours. For instance, case vignettes have been shown to be an ecologically valid tool in educational research on tracking decisions (Krolak-Schwerdt et al., 2018), and estimates from vignettes align with behavioural benchmarks in analyses determining which immigrant attributes generate support for naturalisation (Hainmueller et al., 2015). In an attempt to balance rigour and relevance, Chapter 5 of this thesis therefore contains a vignette study to investigate the motivations that observers attribute to second- and third-party punishers.

## **1.4 Overview of this Thesis**

In this thesis, I aim to go beyond experimental control and isolated processes by investigating reputation in complex real-world environments.

In **Chapter 2** I investigate how reputation management varies across social relationships and situational contexts in everyday life. Participants used a phone app to complete daily assessments for a week, where they reported on recent instances in which they managed their reputation. I found that social relationship factors influenced reputational concern more than social visibility cues.

**Chapter 3** uses a weekly survey study conducted across six weeks to determine the factors that influence a punisher's reputation. I found that the perceived justness of punishing and whether a punisher is perceived to be motivated by altruistic or selfish motives determine reputational consequences, with punishers receiving better evaluations when they were perceived to be acting altruistically. The reputation of punishers was not strongly influenced by the type of punishment (TPP or SPP). These findings highlight the importance of considering the context in which punishment occurs and the motives of the punisher when evaluating the reputational consequences of punishment.

In **Chapter 4** I investigate whether we know how our cooperative and uncooperative behaviours influence our reputation using an adaptation of the Social Accuracy Model (Biesanz, 2010; Lees et al., 2022). I solicited written descriptions of recent cooperative and uncooperative behaviours with accompanying self- and meta-perceptions, followed by actual judgements on the same items. For both cooperative and uncooperative behaviours, actors understood how they were uniquely perceived relative to the average actor and showed insight into how they would be misperceived by observers. At the same time, actors overestimated the transparency of their motives to others. Additionally, when reflecting on their uncooperative actions, actors exhibited a systematic negative bias, expecting others to view them less favourably than they actually did.

In **Chapter 5** I use a vignette study to investigate the motivations that observers attribute to second- and third-party punishers. Using the same methods as Chapter 4, I also examine whether punishers can identify the motives observers ascribe to their actions and whether observers accurately perceive punishers' actual reported motives. I found that third-party punishers received more positive evaluations than second-party punishers. Observers tended to associate TPP with prosocial motivations, while SPP was linked to concerns of self-protection and formidability. Punishers accurately predicted these judgements but still displayed biases.



Finally, in **Chapter 6** I present a registered report containing five economic game experiments investigating whether deliberating over personal cost and target impact signals trustworthiness in punishment and helping contexts. Previous work has shown that uncalculated helpers are perceived as more trustworthy – and I extended this approach to ask whether the same is true for uncalculated punishment. Although I replicated earlier helping results, this did not extend to punishment or target impact deliberations.

## **Chapter 2**

### **Reputation Management In Everyday Life**

## 2.1 Introduction

A person's reputation can serve as a signal that reduces uncertainty about their expected behaviour (Posner, 1997; Zinko et al., 2007). Attending to another's reputation can thereby reduce costs in forming and maintaining social relationships (Garfield et al., 2021). Consequently, individuals monitor others' reputations to inform decisions about when, how, and with whom to engage (Dores Cruz et al., 2021). Given that a positive reputation can enhance success in dating, employment, social support, wellbeing, and reproductive outcomes (Barker et al., 2019; Steinmetz et al., 2016), and conversely, a negative reputation can diminish social standing, decrease chances of being selected as an interaction partner, or result in social exclusion (Dores Cruz et al., 2021), it becomes valuable for individuals to influence how they are perceived by others—essentially, to manage their reputation.

Reputation management is a strategic behaviour aimed at shaping others' perceptions of oneself, driven by concerns about external evaluations (Banerjee, 2002). Strategies include concealing, withholding, or even misrepresenting information, as well as adjusting one's behaviour (e.g., Piazza & Bering, 2010; Watling, 2019; Weiss & Feldman, 2006; Wu et al., 2016). To effectively manage one's reputation, one must be aware of what other people's expectations and values are, and what is considered normative behaviour – which may differ depending on the respective social group or specific situation (Watling, 2019). For instance, one might downplay one's competence when trying to appear warm but downplay one's warmth when wanting to appear competent (Holoien & Fiske, 2013). We thereby form and maintain multiple reputations in our everyday lives, as we are for instance viewed differently by our friends than by people at work (Cavazza et al., 2015), and form reputations across many domains (Macfarlan & Lyle, 2015; Marcfarlan et al., 2013).

However, much of the research on reputation management conducted by psychologists and economists has focused on a narrow range of domains (predominantly prosociality), and has been studied with highly controlled and nearly context-free simulations and experiments (Giardini & Wittek, 2019), often using one-shot or short economic games. These studies are often unlike the situations that we may encounter in our everyday lives in that they (i) frequently require participants to make forced choices which may not adequately capture how they would behave in the real world, and (ii) are mostly based on interactions between strangers, neglecting other common interaction partners in daily life. Reputation management studies from the corporate world have been restricted to businesses and the workplace, again mostly focusing on reputations of capability and prosociality (e.g., Zinko et al., 2007). Sociologists and

anthropologists have worked with a wider range of reputation domains (e.g., also capability (Macfarlan & Lyle, 2015), as well as aggression (Garfield et al., 2019) and sexuality (Abu-Lughod, 1990; Macfarlan & Lyle, 2015). Yet, even in these fields, factors influencing reputation and reputation management have usually been studied in isolation, with a frequent focus on single domains or behaviours at a time (but see Garfield et al., 2021, who identified distinct reputation domains from ethnographic literature), which are often context-dependent and not easily generalisable beyond the specific population being studied. Hence, there has been a call even in anthropology to consider subtle and cumulative everyday behaviours, and to extend studies past isolated, short-term or one-off interactions (Bliege Bird et al., 2018). Although laboratory studies are vital for testing theories, the present study aimed to go beyond ensuring experimental control and investigating processes in isolation, by focusing on humans in their complex real-world environments, and studying the reputation management behaviours of our participants in their natural settings.

Particularly when considering a broader array of reputation domains, little attention has been given to understanding whether or how individuals' concern for reputation varies depending on context features. The following will highlight some factors that may influence reputation concern.

Firstly, people may be more concerned about their reputation when they feel socially visible. Whenever an individual's behaviour is observable, their reputation is at stake. Consequently, people often act more prosocially and take fewer risks in situations where their behaviour can be observed or will be otherwise known about (e.g. Andreoni & Petrie, 2004; Bereczkei et al., 2007; Harbaugh, 1998; Hoffman et al., 1994; Jia & McMahon, 2020; Milinski et al., 2006; Samu et al., 2020; Semmann et al., 2004; Yoeli et al., 2013). For example, Yoeli et al. (2013) found that residents who signed up for an environmental programme using their names saved more energy than those who signed up anonymously. The impact of anonymity was most significant among apartment residents and homeowners, where the potential for repeated interactions was higher, thereby also showing that factors such as social visibility can interact with factors such as likelihood of future interactions (mentioned further below). However, engaging in prosocial behaviours does not always enhance reputation, as observers might suspect that those actions were motivated by strategic gains rather than a genuine cooperative disposition (Raihani & Power, 2021). Thus, sometimes people avoid helping publicly or hide their good deeds (Raihani & Power). Similarly, punishers both advertise (Kurzban et al., 2007; Piazza & Bering, 2008) and conceal (Rockenbach & Milinski, 2011) their punitive actions,

indicating that punishment sometimes does and sometimes does not yield reputation benefits. Thus, although it leads to various reputation management strategies, concern for reputation appears to be increased by audience presence and lack of anonymity (although this concern is not often explicitly measured). Overall, individuals are more likely to engage in reputation management behaviours when they are concerned about their reputation (de Araujo, 2014).

The human development of language, and consequently gossip, has expanded the scope within which our reputations impact us (Garfield et al., 2021). Due to our ability to gossip, we can gain information about other people's past behaviours without having to observe them and use this gossip information – which we overwhelmingly believe (Dores Cruz et al., 2021) - to guide our potential interactions with them (Barclay, 2015; Dunbar, 2006; Smith, 2010). Moreover, as gossip increases the 'audience' that can learn from our actions, the consequences of obtaining a good or bad reputation are enlarged (Nowak & Sigmund, 2005). Recent research, for instance, found that the increased social visibility in rural communities compared to cities led to those individuals experiencing higher levels of reputational concern (Danielson et al., 2023).

A factor related to social visibility and the capacity to be gossiped about is a person's social connectedness, which can amplify the impact that an individual's behaviour may have on their reputation (Cavazza et al., 2014). For instance, during negotiation tasks carried out between students across several weeks, the link between an individual's reputation and their behaviour was stronger for those who were more well-known and who received more attention within that community (Anderson & Shirako, 2008). Individuals also tend to act more prosocially towards other people who are well-connected within their social networks as these people have more gossip capacity (Curry & Dunbar, 2011; Wu et al., 2015). This is especially true when the interaction partner, or those whom they could gossip to, are likely to interact with the individual in future (e.g., Yoeli et al., 2013; Wu et al., 2016). When the likelihood of future interactions is low, the reputational consequences of an individual's actions diminish, as the possibilities to benefit from indirect reciprocity (IR) or reputation-based partner choice (RBPC) are lower to begin with.

Individuals do not indiscriminately try to manage everyone's impression of themselves. Instead, people tend to exhibit more prosocial behaviour towards ingroup members compared to outgroup members (Everett et al., 2015). Even young children demonstrate more cooperative behaviour when they are observed by another child who belongs to their ingroup rather than an outgroup (Engelmann et al., 2013). Furthermore, group entitativity and being accountable to

ones ingroup increases concern for one's reputation, a link that is mediated by fear of social exclusion (Cavazza et al., 2014; Pagliaro et al., 2016).

Reputation management strategies are also likely to vary with individual-level characteristics and traits. It is, for instance, possible that individual differences, such as political ideology, may influence whose view of ourselves we are most concerned about. Recent research has stated that differences in political ideology reflect core psychological differences, and demonstrated that liberals are more compassionate towards more encompassing entities (universalism), caring more about friends relative to family, whilst conservatives care more about less encompassing entities (parochialism) (Waytz et al., 2019). This variation might also influence the contexts in which people manage their reputation, as political conservatives may be more concerned with how they are viewed by family, while political liberals may be more concerned with how they appear to more distant connections.

Reputation concern has also been linked to social anxiety, self-esteem, and self-consciousness (Combs & Penn, 2004). Social anxiety, for instance, influences how individuals manage their reputation through selective disclosure of achievements (Dubey et al., 2024). This behaviour is observable even on social media platforms like Facebook, where the desire to impress peers correlates with heightened levels of social anxiety and self-consciousness (Lee & Jang, 2019). Moreover, individuals who derive their self-esteem from external approval and prioritise avoiding social harm are particularly sensitive to their reputation (Cavazza et al., 2015; Pfattheicher, 2015). These insights underscore the multifaceted nature of reputation management, influenced by both individual psychological dynamics and broader social contexts.

Similar to recent research which has studied social behaviours in the real world (e.g., Columbus et al., 2021; Dores Cruz et al., 2021; Hofmann et al., 2014; Molho et al., 2020; Weiss et al., 2021), the present study used diary methods to test how reputation management varies across social relationships and situational contexts. This method involves asking participants to communicate about the behaviours they engaged in during the day, which reduces recall bias as these self-reports are gathered close in time to when the behaviours occurred (Schwarz, 2012; Scollon et al., 2009), and increases ecological validity as it is conducted in the participants' natural environment, capturing dynamic person-by-situation interactions (Uy et al., 2010). Using a survey app for a week, participants reported daily on their reputation management behaviours. In combination with an intake questionnaire, I aimed to investigate various individual (i.e., pertaining to a participant's attributes or personality) and situational (i.e.,

relating to the context and relationship between a participant and their social partner(s)) aspects of reputation management in real-world settings.

I had two main hypotheses. Firstly, participants would report higher levels of reputational concern when they managed their reputation compared to when they engaged in other social interactions. Secondly, participants would be more concerned about their reputation (i) when interacting with well-connected others, (ii) when interacting with in-group versus out-group members, (iii) when they valued and (iv) felt close to the person(s) whose view of themselves they wanted to manage, (v) when future interactions were likely, and (vi) when they had an audience. Secondary analyses investigated the potential effects of participants' traits on their reputational concerns. Specifically, these analyses examined whether participants with higher levels of social interaction anxiety and social phobia were more concerned about their reputation, and whether political conservatives were more concerned with how they are viewed by family and close friends compared to political liberals.

The preregistration can be found in Appendix A1.

## **2.2 Methods**

### **2.2.1 Participants**

Participants were recruited via Prolific Academic in August 2021, and only those who were at least 18 years old, fluent in English, and possessed a smartphone with either an Android or iOS operating system were invited to take part in the study. Only participants who had joined Prolific at least one week prior to data collection were invited, as there was a reported influx of new Prolific participants due to a viral Tik Tok video which was said to be skewing demographics. 161 participants took part in both the intake survey and the scheduled daily surveys. However, one of these participants was removed, as she told the researcher that she had not completed the scheduled surveys properly, leaving a final sample size of 160. The sample had a mean self-reported age of 25 years ( $SD = 6.28$ , range: 18-51 years), with 102 females, 53 males and 5 individuals identifying as ‘other’. Many participants were highly educated (53.4% had a bachelor’s or higher degree) and were predominantly White (71%). 29 different nationalities took part, with larger portions of participants being South African (20%) and Polish (19%). All participants provided informed consent before participating and the research was approved by the UCL Research Ethics Committee (Project ID: ICN-PWB-NR-23-10-20A).

### **2.2.2 Intake Questionnaire**

First, participants completed an intake questionnaire, which was implemented in Qualtrics. Participants were informed about the study content and requirements, specifically highlighting information about the daily assessment phase and the app to be used. They were told that the scheduled surveys would assess their everyday reputation management, and that they would be asked whether they recently ‘engaged in intentional actions to create or maintain a desired image of [themselves] in others – in other words, whether [they] have influenced the way others perceive [them]’. Participants were moreover instructed to report on a specific instance, and to use initials or pseudonyms if they talked about themselves or others in free text box responses to protect their anonymity. Subsequently, participants were asked to provide demographic information, before completing several trait measures presented in a random sequence. Within these questionnaires all items were provided in randomised order.



All Cronbach's alphas ( $\alpha$ ) reported in this chapter indicate the internal consistency of each scale based on the data collected in this study. Reputation concern was measured with the 7-item Concern for Reputation Scale ( $\alpha = 0.82$ ) (De Cremer & Tyler, 2005) in which items are rated from 1 (*Not at all characteristic of me*) to 5 (*Extremely characteristic of me*). Social and economic conservatism ( $\alpha = 0.78$ ) (SECS; Everett, 2013) was assessed by asking participants to score 12 issues from 0 to 100, in which scores of 0 indicated greater negativity, scores of 100 indicated positivity, and scores of 50 indicated neutral feelings about the issue. The SIAS-6 and SPS-6 ( $\alpha = 0.89$ ) (Peters et al., 2012) were used to assess levels of social interaction anxiety and social phobia. For this, 12 items were measured on a 5-point scale from 0 (*Not at all characteristic or true of me*) to 4 (*Extremely characteristic or true of me*). The Social Perspective Taking Scale ( $\alpha = 0.86$ ) (Gelbach et al., 2008) used 7 items to measure engagement on a 5-point scale from 1 (*Almost Never*) to 5 (*Almost All the Time*). Lastly, the 44-item Big Five Inventory (John & Srivastava, 1999), in which items were rated from 1 (*Strongly disagree*) to 5 (*Strongly agree*), was administered (extroversion:  $\alpha = 0.8$ , agreeableness:  $\alpha = 0.79$ , conscientiousness:  $\alpha = 0.76$ , neuroticism:  $\alpha = 0.79$ , openness:  $\alpha = .075$ ). At the end of the intake, information relevant to the completion of the scheduled surveys was repeated, and participants were given a link to PIEL (Participation In Everyday Life) Survey App Instructions which contained instructions about how to find the PIEL app, load and activate the study survey, as well as how to share their data at the end of the study. To verify that participants successfully set up the daily assessments in the PIEL app, the Prolific advertisement, information sheet, and PIEL app instructions incentivised immediate activation of the scheduled surveys by offering a £0.10 bonus. The bonus was awarded if participants sent a code word within 24 hours of completing the intake questionnaire. This code word was visible on the placeholder screen of the study survey, and therefore only became accessible to participants once they had activated the daily surveys. Median completion time for the intake survey was 14 minutes.

### **2.2.3 Daily Assessment Phase**

The one-week daily assessment phase began either on the day or the day after the survey was activated in the PIEL app, depending on whether the survey was activated before or after 19:00, which was the scheduled time for the daily surveys. The daily surveys were run in the PIEL app to more easily record multiple observations from the same participants over a seven-day period. The PIEL app is free and available to both Android and iOS users. Unlike many platforms, no third-party server or database is required as data is saved on participants' own devices before

being emailed to the researcher by the participant at the end of the study. Once participants activated the study, a placeholder screen showed them the study's start and end date, as well as the number and percentage of surveys which they had completed. Participants received daily in-app notifications to complete the study at 19:00 (their own time) and had 2 hours (until 21:00 their time) to respond to the survey before it was recorded as missed. I chose to schedule the surveys once a day based on my piloted frequency of which people can recall having managed their reputation<sup>1</sup>. The response window was also based on piloting, providing enough time for busy individuals to complete the survey. Scheduling it in the evening allowed participants to have experienced social interactions, increasing the likelihood of having managed their reputations. Whenever the app was not opened, further in-app reminder notifications were sent after (i) 30 minutes, (ii) 1 hour and (iii) 1 hour 30 minutes to reduce any memory burden and to increase compliance. Following completion of the final scheduled survey, the app prompted participants to send their recorded responses to the researchers, simply by selecting "OK". The median completion time for scheduled surveys was 3 minutes, with an average response rate of 89%. As per my preregistration, I used free text box responses to exclude responses. Thirty-six responses were removed because participants indicated that they had not had any social interactions with anyone that day and therefore had nothing to report (e.g., "I did not have any social interactions today"). Another response was removed as it included three separate incidents rather than one specific event, and three participants began another scheduled survey after their 7-day period ended, for which the further responses were not considered. In total, 994 scheduled survey responses were included in the analysis, in which participants reportedly managed their reputation in 57% of cases ( $k = 571$ ) and answered questions about their recent social interactions in the remaining instances ( $k = 423$ ).

## 2.2.4 Survey Content

As a reminder, participants read a brief definition of reputation management ("Reputation management' is when we engage in actions that will create or maintain a desired image of ourselves in others") at the beginning of each scheduled survey and were asked to try to recall

---

<sup>1</sup> Piloting involved testing the survey with around 60 individuals, including friends, family members, neighbours, and friends of my parents, who were of various ages, nationalities, and life situations (e.g., homemakers, retirees, working professionals, and students). Participants provided feedback on topics such as how frequently they recalled managing their reputation, their preferred time windows for survey completion, and their tolerance for reminders (i.e., what they found helpful versus annoying). This feedback was used to refine the survey frequency, response window, and timing to suit participants' schedules and optimise recall accuracy.

a recent instance in which they managed their own reputation. As a tip, they were told that it may help to think about their recent social interactions. Subsequently, they were asked whether they had managed their reputation since the last survey. Participants could reply either “yes” or “no” each time. Whenever participants indicated that they had managed their reputation, they were asked to give both a brief description of *what* happened (i.e., information about what they did and what the situation was), and *why* they managed their reputation (i.e., what their motivations and feelings were regarding how they wanted to be perceived). Due to their rich content, these free text box entries could simultaneously act as attention, comprehension, bot, and English proficiency checks. I furthermore measured reputational concern (“How concerned did you feel for how you might appear to the other person(s)?”) on a 0.01-increment slider scale from 0 (*Not at all concerned*) to 1 (*Extremely concerned*). I then asked additional questions about the person(s) whose perception of themselves the participants wanted to manage. This included questions about relationship value (“How much do you value the relationship with the person(s) whose perception you wanted to manage?”), measured on the same 0.01-increment slider scale from 0 (*Not at all*) to 1 (*Very much*), and feelings of relationship closeness (“At this moment, how close do you feel to the person(s) whose view of yourself you wanted to influence?”; 0 = *Not at all close*, 1 = *Very close*). As a proxy for determining whether others were more likely to belong to a participant’s in-group or out-group, I also measured the relationship participants had with the person(s) (i.e., stranger, acquaintance, colleague/classmate, instructor/superior, friend, romantic partner, family, or other (to be specified by the participant)). Following this, I measured the perceived connectivity of the other person(s) (“How well-connected is the person(s) to others who know you (i.e., do you know many of the same people)?”; 0 = *Not at all connected*, 1 = *Very connected*). The likelihood of future interactions taking place between the participant and the person whose perception of themselves they wanted to manage (“How likely is it that you will interact with the person(s) in future?”) was also assessed on a slider scale from 0 (*Not at all likely*) to 1 (*Very likely*). Lastly, participants reported whether they managed their reputation with or without an audience present (“Did you manage your reputation while you were alone and unobserved, or in front of others (this includes the person(s) whose perception you were managing)?”). Whenever participants indicated they had not managed their reputation, they answered equivalent questions about their most recent social interaction instead. This condition had one less question, as the last question about audience presence was not included. This alternative condition was both included to prevent participants from simply replying “no” to purposely cut the survey short, and more importantly to act as a reference point to be used in the analysis of my first hypothesis.

Refer to Appendix A2 for study materials (including the Qualtrics intake survey, daily survey file, and PIEL survey instructions). For a random sample of participant responses showcasing the diverse range of behaviours and desired reputations reported, see Appendix A3.

### **2.2.5 Participant Payment**

Participants received £2 for completing the intake questionnaire. However, they were told that they also had to take part in the daily assessment phase, by completing at least three scheduled surveys, to be compensated. As mentioned above, participants were incentivised with a £0.10 bonus to activate the survey file within 24 hours, which they demonstrated by sending a codeword displayed on the app's placeholder screen. Participants were given £0.45 for each completed scheduled survey (max. £3.15). They could earn an additional £1 bonus for completing at least 80% of the surveys (6/7). In total, participants could therefore earn a maximum of £2 (intake) + £0.10 (prompt activation) + £3.15 (scheduled surveys) + £1 (80% survey bonus) = £6.25. On average, participants received £5.85 ( $SD = 0.75$ ).

### **2.2.6 Analysis Plan**

Mixed-effects models in this report were conducted in R (R Core Team, 2019) using the packages 'lme4' (Bates et al., 2015), 'lmerTest' (Kuznetsova, 2017), 'r2glmm' (Jaeger, 2017), and 'emmeans' (Lenth, 2019). All reported coefficients are unstandardised.

To assess whether participants were more concerned for their reputation when they were managing their reputation compared to when they were engaging in other social interactions, a linear mixed effects model was fit using Restricted Maximum Likelihood (REML) criterion. *T*-tests used Satterthwaite's method. The model included participant ID as a random effect with random intercepts (formula:  $\text{concern} \sim \text{condition} + (1|\text{ID})$ ).

Throughout this thesis I frequently chose to use mixed effects models because they allowed me to account for both fixed and random effects in the data (Bates, 2010). This approach is particularly useful in repeated-measures designs, where observations are not independent. In this study, each participant provided multiple responses, as they were presented with the same survey daily. Mixed effects modelling is especially well-suited to repeated-measures data because it allows for correlations between multiple observations from the same individual

(Oberg & Mahoney, 2007). By estimating covariance parameters, mixed effects models capture these correlations, improving the accuracy of the estimates (West et al., 2022).

Additionally, mixed effects models offer flexibility in handling missing data, allowing participants with missing responses to remain in the analysis. In contrast, approaches such as repeated-measures Analysis of Variance (ANOVA) typically exclude entire participants with incomplete data from the analysis, which can reduce the study's statistical power (Wainwright et al., 2007; Little & Rubin, 2019). In mixed models each observation represents just one of many responses from an individual, meaning that the removal of a single observation has a smaller effect on the overall analysis compared to ANOVA, which often treats all responses as part of the same observation (Brown, 2021). Mixed-effects modelling is therefore considered a useful analytic tool that is often suitable for situations where the use of e.g. standard ANOVAs, repeated measures ANOVAs, or multiple regression may be less appropriate (Brown, 2021).

To examine whether participants reported higher levels of reputational concern under specific conditions, I employed two distinct analysis methods to address the same issue: one that was preregistered (six individual models) and another that offers a more holistic analysis (a singular model). The preregistered method is presented first, followed by the additional, arguably more appropriate, method.

To test whether participants reported higher levels of reputational concern (i) when interacting with well-connected others, (ii) when interacting with individuals who are in-group rather than out-group members (using relationship type as a proxy), (iii) when they valued and (iv) felt close to the person(s) whose view of themselves they want to manage, (v) when future interactions were more likely, and (vi) when they had an audience, six individual models (formula: concern ~ variable of interest + (1|ID)) were fit using a subset of the data set, including only the cases in which participants managed their reputation ( $k = 571$ ). All mixed models controlled for the random effect of participant ID. As six individual hypotheses were tested, each with the potential to contribute to Type I errors (false positives), a Bonferroni correction was applied to account for multiple hypothesis testing, resulting in an adjusted significance threshold of  $\alpha = 0.0083$  ( $0.05/6$ ).

Recognising the potential limitations of examining each predictor in isolation, a second method was employed to fit these factors into a singular, comprehensive model.

Given the high correlation between relationship closeness and relationship value ( $r = .83$ ), one of these variables was removed from the full model to avoid multicollinearity. An ANOVA was

conducted to compare the fit of two global models: one excluding relationship closeness (model 1) and the other excluding relationship value (model 2). The model excluding relationship closeness (model 1) demonstrated a better fit, indicated by lower Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values (model 1: AIC = -37.50, BIC = 23.37; model 2: AIC = -10.49, BIC = 50.38).

Based on these results, I specified a single linear mixed effects model by incorporating all relevant fixed effects - except for relationship closeness - with participant ID included as a random effect. In this holistic model, each predictor's effect represents its unique contribution to reputational concern, accounting for the variance explained by other predictors (Judd et al., 2017). This differs from the individual models, where each predictor's effect reflects only its bivariate relationship with reputation concern. By controlling for shared variance, the holistic model isolates the unique effects of each predictor, which can result in differences in effect size and direction compared to the individual models (Judd et al., 2017).

Lastly, to examine the relationship between participants' mean concern for reputation scores and other trait scores collected in the intake questionnaire, a Bonferroni adjusted alpha level of  $\alpha = .0063$  (.05/8) was applied. This adjustment was used to determine significance of correlations, accounting for multiple hypothesis testing.

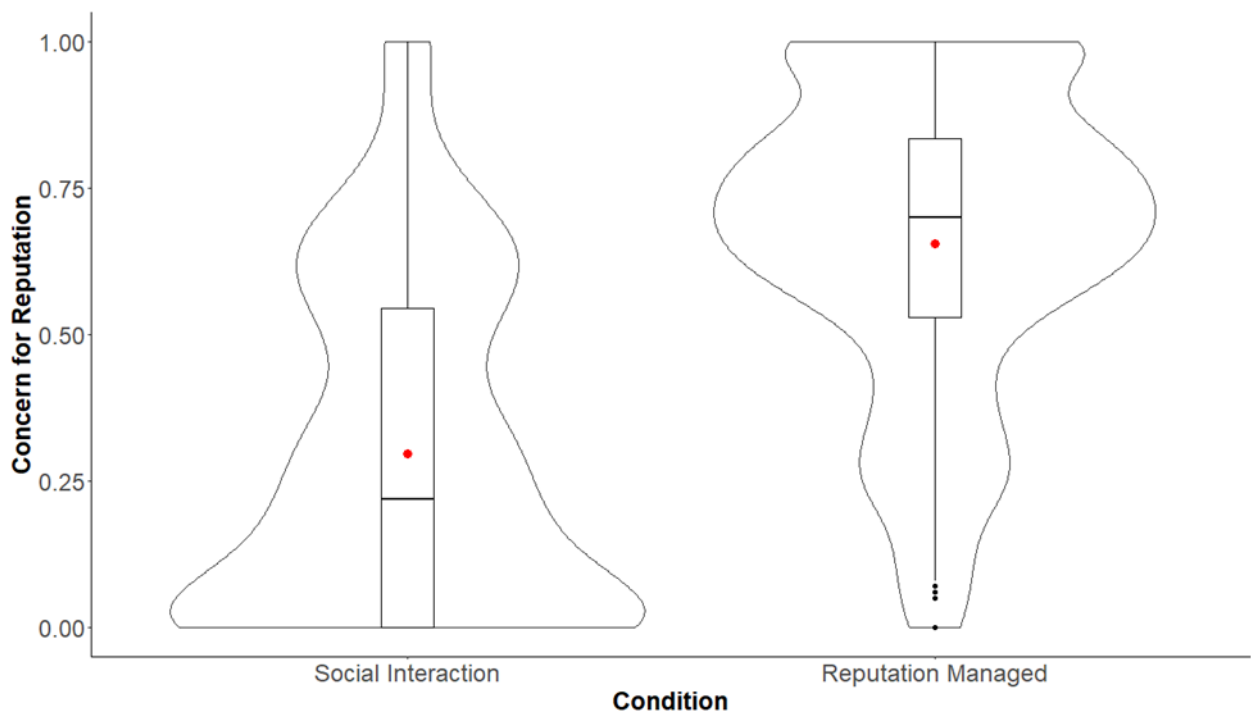
## 2.3 Results

### 2.3.1 Reputational Concern Across Reputation Management and Social Interaction Events

Participants were more than twice as concerned for their reputation during situations in which they were managing their reputation ( $M = 0.66$ ,  $SD = 0.26$ ) than during other social interactions ( $M = 0.30$ ,  $SD = 0.29$ ),  $b = 0.34$ ,  $t(991.9) = 19.65$ ,  $p < .0001$ ; 95% CI [0.31, 0.38] (Figure 2.1). The random effect of participant explained 25% of the residual variance (Marginal  $R^2 = 0.28$ ; Conditional  $R^2 = 0.46$ ). A likelihood-ratio test indicated that including condition improved model fit,  $\chi^2(1) = 326$ ,  $p < .0001$ , which was further supported by lower AIC and BIC values.

**Figure 2.1**

*Levels of reputational concern in situations without and with reputation management ( $k = 571$ )*



*Note:* The violin plots show the distribution density of the data. The black horizontal lines indicate the median values. The black box in the boxplot represents the interquartile range (IQR), with whiskers extending to the furthest points within 1.5 times the IQR, and points beyond this range are considered outliers. Red dots signify mean values.

### 2.3.2 Preregistered Analysis: Individual Factors Predicting Levels of Reputational Concern

First, each independent factor (social connectedness, relationship type, relationship value, relationship closeness, likelihood of future interaction, and audience presence) was considered individually, as preregistered. See Table 2.1 for both the individual models and the holistic model results. The continuous predictors likelihood of future interactions ( $b = 0.15$ ,  $t(554.04) = 4.83$ ,  $p < .001$ , 95% CI [0.09, 0.22]), social connectedness ( $b = 0.11$ ,  $t(563.04) = 3.62$ ,  $p < .001$ , 95% CI [0.05, 0.17]), relationship closeness ( $b = 0.12$ ,  $t(566.62) = 3.82$ ,  $p < .001$ , 95% CI [0.06, 0.18]), and relationship value ( $b = 0.21$ ,  $t(561.10) = 6.94$ ,  $p < .001$ , 95% CI [0.15, 0.27]) were all significant positive predictors of reputation concern (Figure 2.2), with likelihood-ratio tests indicating that an inclusion of predictors would improve model fit (likelihood of future interactions:  $\chi^2(1) = 22.96$ ,  $p < .001$ ; social connectedness:  $\chi^2(1) = 12.95$ ,  $p < .001$ ; relationship closeness:  $\chi^2(1) = 14.49$ ,  $p < .001$ ; relationship value  $\chi^2(1) = 46.39$ ,  $p < .001$ ). The random effect of participant explained a further 32%-33% of the residual variance within each model.

**Table 2.1**

*Effects of Predictors on Level of Reputational Concern: Individual and Holistic Models*

| Predictors                    | Level of Reputational Concern |              |          |                |                |               |          |                |
|-------------------------------|-------------------------------|--------------|----------|----------------|----------------|---------------|----------|----------------|
|                               | Individual Models             |              |          |                | Holistic Model |               |          |                |
|                               | <i>b</i>                      | 95% CI       | <i>t</i> | <i>p</i>       | <i>b</i>       | 95% CI        | <i>t</i> | <i>p</i>       |
| (Intercept)                   |                               |              |          |                | 0.51           | 0.45 – 0.58   | 14.91    | < <b>0.001</b> |
| Relationship Value            | 0.21                          | 0.15 - 0.27  | 6.94     | < <b>0.001</b> | 0.26           | 0.16 – 0.36   | 5.31     | < <b>0.001</b> |
| Future Interaction Likelihood | 0.15                          | 0.09 - 0.22  | 4.83     | < <b>0.001</b> | 0.02           | -0.07 – 0.12  | 0.48     | 0.63           |
| Social Connectedness          | 0.11                          | 0.05 - 0.17  | 3.62     | < <b>0.001</b> | 0.02           | -0.05 – 0.10  | 0.62     | 0.54           |
| Relationship Closeness        | 0.12                          | 0.06 - 0.18  | 3.82     | < <b>0.001</b> |                |               |          |                |
| Audience Presence             | -0.02                         | -0.06 . 0.03 | -0.74    | 0.46           | 0.001          | -0.04 – 0.05  | 0.05     | 0.96           |
| Acquaintance                  | 0.04                          | -0.03 - 0.12 | 1.09     | 0.28           | -0.03          | -0.12 – 0.05  | -0.80    | 0.42           |
| Colleague                     | 0.10                          | 0.03 - 0.17  | 2.66     | 0.008          | -0.01          | -0.10 – 0.07  | -0.34    | 0.73           |
| Superior                      | 0.16                          | 0.07 - 0.24  | 3.67     | < <b>0.001</b> | 0.04           | -0.06 – 0.13  | 0.75     | 0.45           |
| Friend                        | 0.17                          | -0.02 - 0.11 | 1.37     | 0.17           | -0.13          | -0.22 – -0.04 | -2.81    | <b>0.005</b>   |
| Romantic Partner              | 0.17                          | 0.09 - 0.25  | 4.03     | < <b>0.001</b> | -0.05          | -0.16 – 0.06  | -0.85    | 0.40           |
| Family                        | 0.12                          | 0.05 - 0.19  | 3.41     | < <b>0.001</b> | -0.09          | -0.19 – 0.01  | -1.68    | 0.09           |
| Other Relation                | 0.05                          | -0.05 - 0.14 | 0.99     | 0.32           | -0.03          | -0.13 – 0.07  | -0.52    | 0.60           |

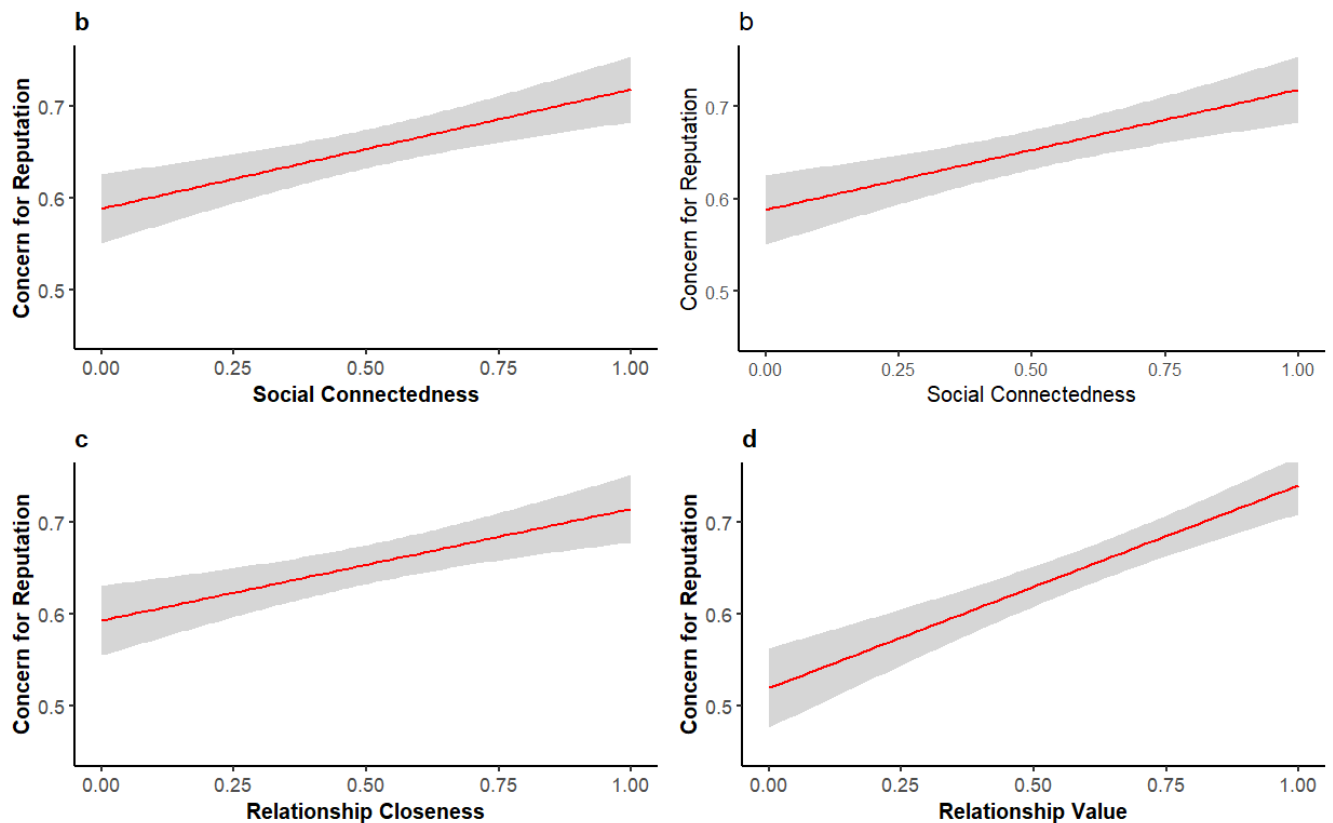
*Note.* This table presents the effects of predictors on reputational concern as assessed by individual and holistic models. In the individual models, each predictor is assessed in isolation, showing its direct relationship with reputational concern. However, this approach does not



adjust for potential confounding effects that may arise when predictors co-vary. By contrast, the holistic model includes multiple predictors simultaneously, with each predictor’s effect representing its “unique effect”—capturing only the unique variance in reputational concern that is not explained by other predictors (Judd et al., 2017). As a result, differences in the significance and magnitude of effects appear between the two approaches. The intercept is not reported in the ‘Individual Models’ section because each model had its own intercept, unlike the ‘Holistic Model.’ Relationship closeness was excluded from the holistic model due to its high correlation with relationship value, as previously described. Significant results are bolded. The individual models have an adjusted significance threshold of  $\alpha = 0.0083$  (0.05/6). The reference level for relationship type (including acquaintance, colleague, superior, friend, romantic partner, family, and ‘other’) is ‘stranger,’ and the reference level for audience presence is ‘audience absent.’

**Figure 2.2**

*Associations of reputational concern with (a) likelihood of future interactions, (b) social connectedness, (c) relationship closeness, and (d) relationship value.*

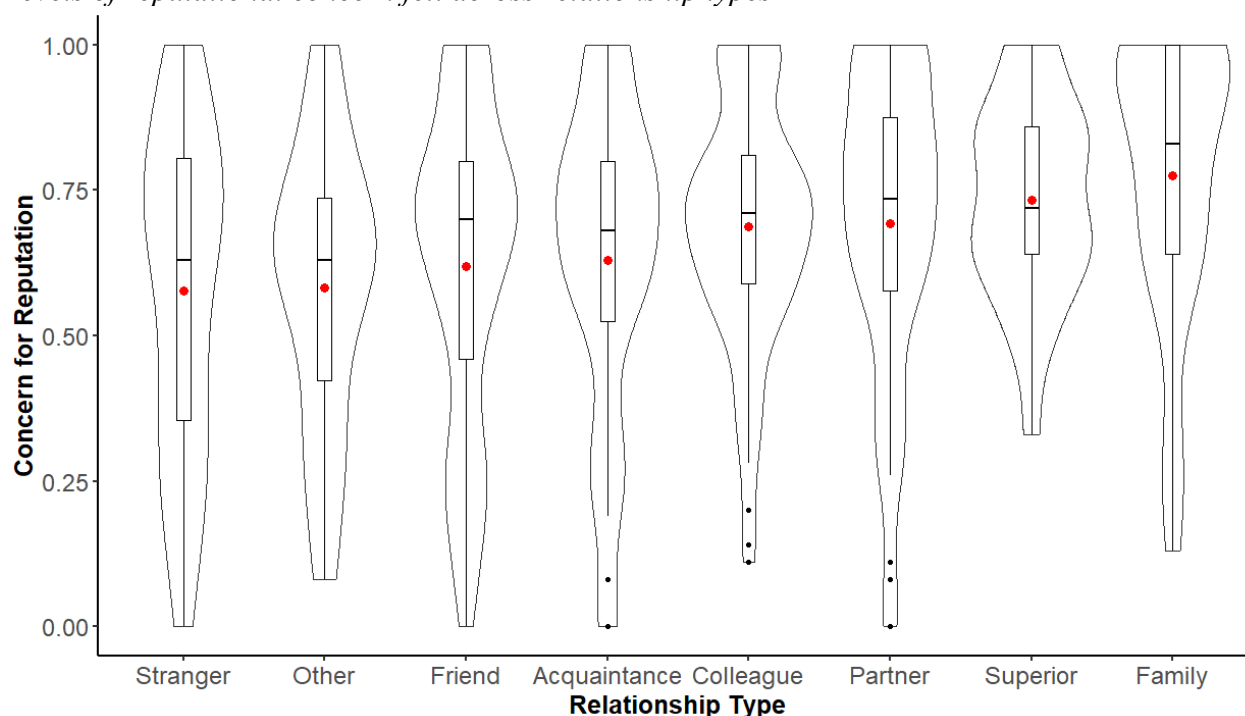


*Note.* This figure displays the relationship between concern for reputation and (a) likelihood of future interactions, (b) social connectedness, (c) relationship closeness, and (d) relationship value. The red lines represent regression lines, indicating the trend of the data, while the shaded area around those lines show the confidence interval of the regression, providing an estimate of the uncertainty around the trend line.

When examining participants' reputational concern across various relationship types, I found that participants most frequently reported wanting to influence the impression their friends (21%) and strangers (17%) have of them. Following that, they most often reported managing their reputation in front of family members (15%), colleagues or classmates (14%), acquaintances (10%), their romantic partners (10%), superiors (9%), and others (5%; mostly neighbours or roommates). The linear mixed model predicting levels of reputational concern felt across different relationship types showed the model's intercept, corresponding to relationship = stranger, was at  $b = 0.58$  ( $t(561) = 21.48$ ,  $p < .001$ , 95% CI [0.53, 0.63]). Participants reported the least amount of concern for strangers, their most distant connections. A Bonferroni correction was applied to adjust for multiple comparisons across relationship types, with an adjusted significance level of  $\alpha = .0063$  (i.e.,  $.05/8$ ). The relationship effects of romantic partner ( $b = 0.17$ ,  $t(551.54) = 4.03$ ,  $p < .001$ , 95% CI [0.09, 0.25]), superior ( $b = 0.16$ ,  $t(557.39) = 3.67$ ,  $p < .001$ , 95% CI [0.07, 0.24]), and family member ( $b = 0.12$ ,  $t(559.77) = 3.41$ ,  $p < .001$ , 95% CI [0.05, 0.19]), were statistically significantly different to the effects of stranger, while participants did not report significantly different levels of reputation concern for colleague ( $b = 0.10$ ,  $t(537.9) = 2.66$ ,  $p = .008$ , 95% CI [0.03, 0.17]), friends ( $b = 0.17$ ,  $t(545.32) = 1.37$ ,  $p = .17$ , 95% CI [-0.02, 0.11]), acquaintances ( $b = 0.04$ ,  $t(543.79) = 1.09$ ,  $p = .28$ , 95% CI [-0.03, 0.12]), and “other” relationship types ( $b = 0.05$ ,  $t(518.97) = 0.99$ ,  $p = .32$ , 95% CI [-0.05, 0.14]) compared to strangers. The random effect of participant explained a further 32% of the residual variance, and a likelihood-ratio test indicated that including relationship type improved the model fit,  $\chi^2(1) = 28.60$ ,  $p < .001$ . Figure 2.3 visualises that participants were most concerned about their reputations when influencing how they may be viewed by their romantic partners ( $M = 0.78$ ,  $SD = 0.25$ ), followed by concern for how their superiors ( $M = 0.73$ ,  $SD = 0.17$ ), family members ( $M = 0.69$ ,  $SD = 0.26$ ), colleagues ( $M = 0.69$ ,  $SD = 0.21$ ), acquaintances ( $M = 0.63$ ,  $SD = 0.27$ ), friends ( $M = 0.62$ ,  $SD = 0.27$ ), other relationship types ( $M = 0.58$ ,  $SD = 0.25$ ), and least of all strangers ( $M = 0.58$ ,  $SD = 0.28$ ), may judge them.

**Figure 2.3**

*Levels of reputational concern felt across relationship types*

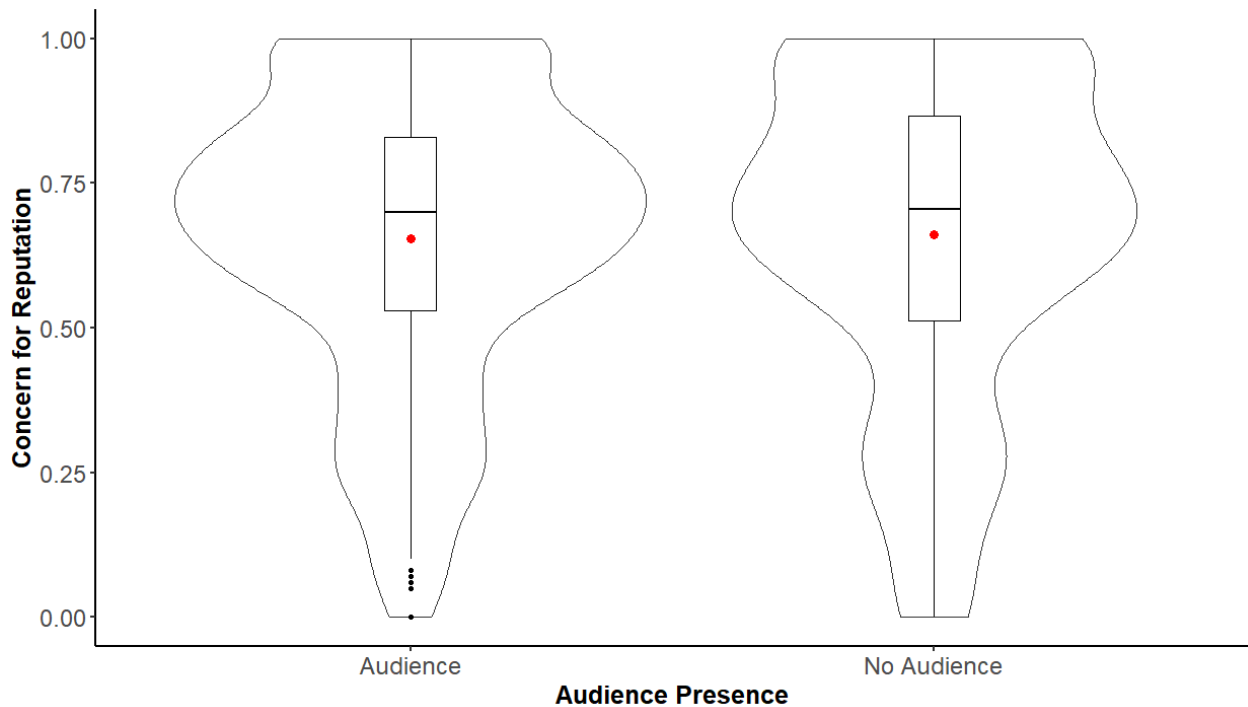


*Note:* The violin plots show the distribution density of the raw data. The black horizontal lines indicate the median values. The black box in the boxplot represents the IQR, with whiskers extending to the furthest points within 1.5 times the IQR, and points beyond this range are considered outliers. Red dots signify mean values. Higher values indicate higher concern for reputation.

Audience presence did not significantly predict levels of reputational concern ( $b = -0.02$ ,  $t(565.89) = -0.74$ ,  $p = .46$ , 95% CI  $[-0.06, 0.03]$ ), as participants reported similar levels of concern for situations in which an audience was present ( $M = 0.65$ ,  $SD = 0.25$ ) versus absent ( $M = 0.66$ ,  $SD = 0.28$ ) (Figure 2.4). Overall, participants reportedly managed their reputation in front of others 72% of the time, and tended to consider themselves as engaging in reputation management whilst ‘alone’ when nobody else was physically present, such as when they were on the phone or online, or when nobody was present whilst the reputation management behaviour itself was occurring, although it would later be seen by others (e.g., ‘I cleaned the house because we had guests coming over. I didn’t want them to think we are untidy’).

**Figure 2.4**

*Concern for reputation depending on audience presence*



*Note:* The violin plots show the distribution density of the raw data. The black horizontal lines indicate the median values. The black box in the boxplot represents the IQR, with whiskers extending to the furthest points within 1.5 times the IQR, and points beyond this range are considered outliers. Red dots signify mean values.

### **2.3.3 Additional Holistic Analysis: Combined Predictors of Reputational Concern**

When considering the independent variables together, only the predictors relationship value and relationship type were still significantly associated with concern for reputation (Table 2.1). With all factors taken into consideration at once, the model output indicated that reputation concern was significantly lower for friends compared to strangers, and that participants were directionally only more concerned about how their superiors compared to strangers may view them.

### 2.3.4 The Influence of Traits on Concern for Reputation

Lastly, the relationship between participants' mean concern for reputation scores and other trait scores collected in the intake questionnaire were examined (adjusted alpha level = .0063). Only agreeableness ( $p = .0037$ ) (Figure 2.5) and social interaction anxiety and social phobia ( $p = .0036$ ) (Figure 2.6) were significantly correlated with concern for reputation scores, although these positive associations were weak ( $r = 0.23$  for both). Extroversion ( $r(158) = 0.16, p = .04$ ), conscientiousness ( $r(158) = 0.18, p = .022$ ), neuroticism ( $r(158) = 0.14, p = .07$ ), openness ( $r(158) = 0.02, p = .83$ ), social perspective taking propensity ( $r(158) = 0.14, p = .08$ ) and SECS scores ( $r(158) = 0.15, p = .06$ ) were not significantly related to concern for reputation. Degree of political conservatism also did not differentially affect levels of reputational concern felt across relationship types (all  $p > .05$ ).

**Figure 2.5**

*Association between concern for reputation and agreeableness scores*

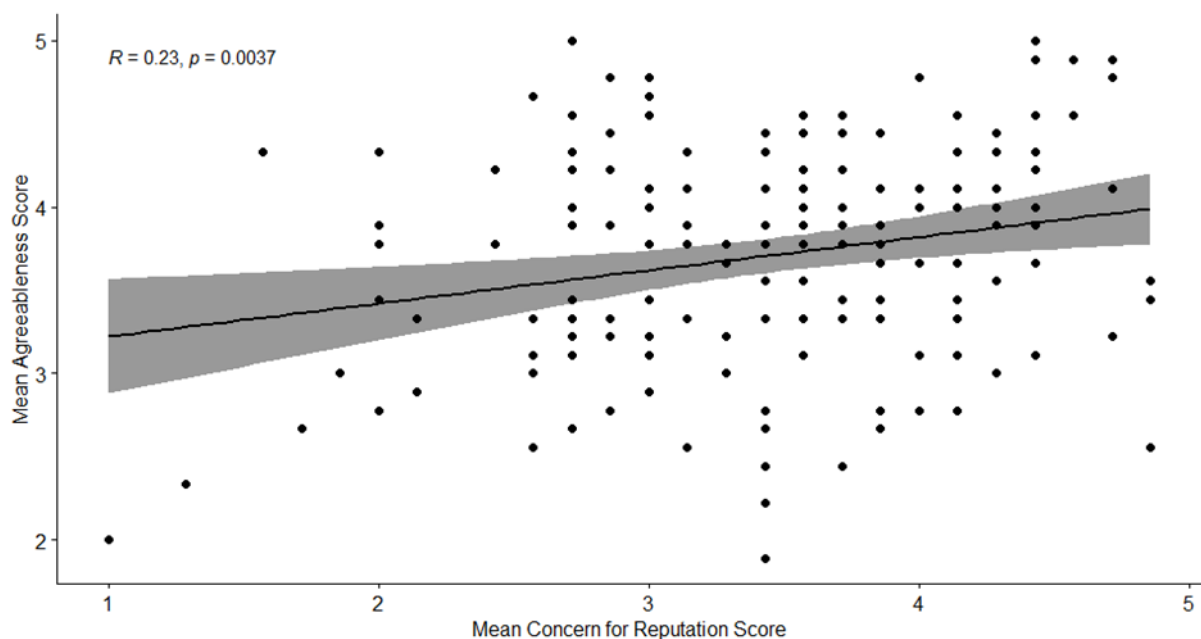
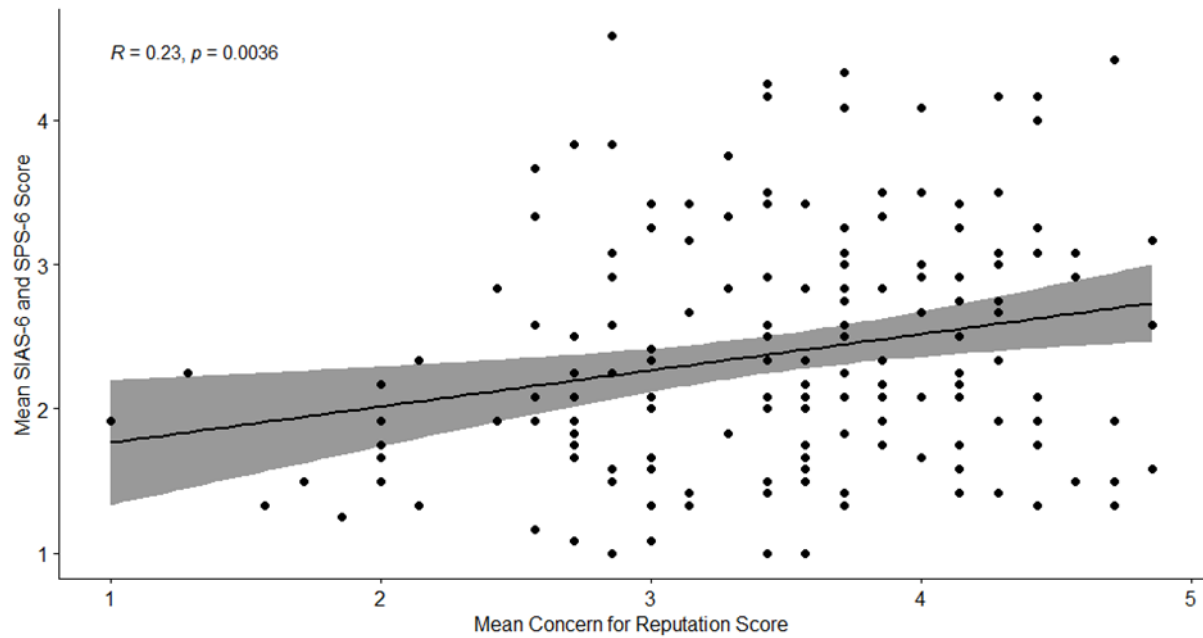


**Figure 2.6**

*Association between concern for reputation and social interaction anxiety and social phobia (SIAS-6 and SPS-6) scores*



## 2.4 Discussion

Prior research on reputation management has typically only considered single reputation domains or behaviours at a time, and has frequently been conducted in highly controlled and nearly context-free ways (Garfield et al., 2021; Giardini & Wittek, 2019). I therefore used diary methods to collect naturally occurring instances of reputation management in participants everyday lives to investigate two main hypotheses. Firstly, whether participants report higher levels of reputational concern during situations in which they manage their reputation compared to other social interactions; and secondly, whether they are more concerned about their reputation (i) when interacting with well-connected others, (ii) when interacting with in-group versus out-group members, (iii) when they value and (iv) feel close to the person(s) whose view of themselves they want to manage, (v) when future interactions are likely, and (vi) when they have an audience.

In line with predictions, participants felt more than twice as concerned for their reputation during situations in which they were managing their reputation than they did during other social interactions. It is possible that participants felt more reputational concern during reputation management events because experiencing higher levels of reputational concern signals or motivates us to manage our reputations, which were the situations participants reported on. This would be in line with the previous finding that the more a person is concerned about their reputation, the more likely they are to engage in behaviours to manage their reputation (de Araujo, 2014).

In support of the second hypothesis, all but one of the predictor variables (audience presence) significantly influenced participants levels of reputational concern, when these factors were considered individually. I found no evidence to suggest that audience presence influenced reputation concern, as participants reported similar levels of concern for situations in which they indicated to have managed their reputation in front of someone else and when they considered themselves alone. However, this does not necessarily contradict prior findings. In the present study, participants generally indicated that they were managing their reputation without an audience when they were physically alone (e.g., interacting with another person over the phone or online), or when they were not interacting with anyone during their reputation management behaviour but the behaviour or character trait would eventually be revealed to others (e.g., cleaning the house before guests arrive). In all reported instances, there was at least a potential that others would come to know about the participant's decision eventually,

indicating that there were no truly private events in this study. It is therefore important to distinguish between immediate audience presence and ultimate anonymity of behaviours.

Conversely, when the predictors were included together in the holistic model, the effects and their interpretation changed. In a multiple-predictor model, each predictor's effect represents its "unique contribution" to reputational concern, capturing only the variance that is not explained by other predictors in the model (Judd et al., 2017). Consequently, when predictors co-vary and are partially redundant in explaining variation in reputational concern, the unique effects observed in the holistic model differ from the bivariate relationships seen in the individual models.

How much the participant valued the person(s) whose view of themselves they wanted to manage appeared to influence reputation concern strongly, both when considered individually and when combined with the other factors. Relationship value in itself has not been extensively studied in relation to reputation management, however, we do try to impress others more if we consider them important (Lee & Jang, 2019). Future research could further study the influence of relationship value on concern for reputation, and also disentangle whether relationships with others have to be valued in a certain way to affect concern. Similarly, I expected participants to report higher levels of reputational concern when they felt close to the other person(s) (Cavazza et al., 2014), which was indeed found when the effect of relationship closeness was considered on its own. Future work should look deeper into the relationship between these factors, for instance by directly manipulating certain factors whilst holding the others constant.

Relatedly, previous research has found that people act more cooperatively with in-group members (Engelmann et al., 2013; Everett et al., 2015). In line with this, when considering the effect of relationship type on reputational concern in isolation, I found that participants felt most concerned about the way their romantic partners would view them, and the least concern for how strangers may think of them. Superiors elicited the second highest amount of reputation concern in participants, which was followed by family, colleagues or classmates, acquaintances, and friends. That participants were more concerned about their reputation in front of acquaintances than friends is less in line with the literature and my predictions, as friends ought to belong more strongly to our in-group. Results diverged even more strongly from prior expectations when the effect of relationship type was considered with the other predictors in one model. Then, participants were least concerned about the way their friends and family members viewed them, and most concerned about how their superiors and strangers might think of them.



Levels of reputation concern directed at superiors was consistently high, and although one's superior may not be considered as an in-group member, it is unsurprising that individuals would want their superiors to think well of them. Here, a good reputation can be directly linked to material benefits such as increased pay or promotions (or e.g., the avoidance of redundancy). Therefore, being favourably evaluated by one's superior is more valuable than a favourable evaluation by someone who has less influence over one's situation (Schlenker & Leary, 1982). Similarly, social network analysis of cooperation among men in a preindustrial society has shown that men gain status by cooperating with higher status individuals (von Rueden et al., 2019), and status considerations explained differences in parental acceptance of their children's close contacts with different outgroups (Munniksma et al., 2012).

However, strangers only elicited high reputation concern in the holistic model; in the individual model, they elicited the least concern. One reason for this difference may stem from partial redundancy among predictors, which affects the unique variance captured in a model with multiple predictors. In the holistic model, relationship type shows its effect on reputational concern after controlling for other factors such as relationship value and social connectedness. Perhaps strangers exerted a stronger influence on reputation concern in the holistic model because this relationship type has the least redundancy; strangers are typically not close, highly valued, or likely to be encountered again.

The finding that participants experienced comparatively high reputational concern for acquaintances and coworkers compared to friends in the individual model (and for strangers compared to friends in the holistic model) may be a result of higher familiarity with close others reducing worry about the unknown. Indeed, Macfarlan et al. (2013) showed that although prosocial reputations track cooperative behaviours, they are anchored across time. Their observations of an Afro-Caribbean village revealed that mild defections are tolerated, and only those who consistently fail to reciprocate acquire negative reputations. Furthermore, many reputations are built cumulatively and slowly with many small acts (Bliege Bird et al., 2018). As our participants will already have built a relationship and reputation with their friends and family, each behaviour would be only one piece of a bigger picture, whereas it would form the entirety of a stranger's opinion. Moreover, if one has already acquired a certain reputation with someone, e.g., for being an intelligent person, presenting more proof of this would not further improve this reputation. Presenting such information to an acquaintance or stranger would have a larger impact on their opinion formation. Besides familiarity measures, such as exploring the age of a relationship, future studies could delve deeper into the influence of relationships on

reputation concern by including additional socio-ecological aspects related to network or population density and relational mobility, which may affect interdependence in social interactions (Columbus et al., 2021; Romano et al., 2021).

The factors social connectedness (i.e., how well connected the other person was to others who may know the participant) and likelihood of future interactions are related to social visibility, or the potential to be gossiped about. In line with previous findings, I expected concern for reputation to increase with these factors, as they increase the ‘audience’ which may learn from an individual’s behaviour, thereby increasing the impact an action can have on their reputation (Cavazza et al., 2014; Curry & Dunbar, 2011; Yoeli et al., 2013; Wu et al., 2015). Although both social connectedness and likelihood of future interactions significantly influenced reputational concern in individual analyses, these predictors were no longer statistically significant in the holistic model. This change likely reflects shared variance with other predictors, such as relationship type and relationship value. These predictors likely overlap in their implications for reputational concern, for instance because high-value relationships often imply a degree of connectedness and high potential for future interactions. By accounting for shared variance, the holistic model isolates each predictor’s unique effect, reducing the apparent influence of predictors that convey similar social functions.

Although social visibility may not be a unique determinant of reputation concerns, factors seemingly associated with it, such as social connectedness in terms of network centrality, could still hold distinct significance in certain contexts. For instance, in small rural communities, the interconnected nature of social networks creates higher social visibility compared to large urban cities. This increased visibility makes individuals in rural areas perceive their reputations as more vulnerable, thereby heightening their concern for maintaining a good reputation. Research has shown that imagining life in a small rural community, as opposed to a large urban one, leads people to express greater concern for their reputations and a stronger reluctance to engage in behaviours that might damage their reputation (Danielson et al., 2023). Moreover, a cross-sectional analysis of the World Values Survey demonstrated a significant correlation between living in smaller communities and a heightened concern for one’s reputation (Danielson et al., 2023). Furthermore, a study examining the support ties of adults in Tamil villages revealed that a reputation for influence had a limited effect on prosocial returns, whereas a reputation for generosity had the strongest correlation with accessing support (Power & Ready, 2018). Notably, this association was partly attributed to the robust effects of reciprocity and shared connections: individuals with generous reputations were more likely to be solicited for

assistance and consequently more likely to receive it. Further, a reputation for generosity was more closely linked to occupying an influential position within one's network (Power & Ready, 2018). As discussed in the general introduction, being well-known and well-regarded is likely advantageous, as it can facilitate access to a broader array of resources and support, as well as the ability to exert influence over others (e.g., Bliege Bird & Power, 2015; Gurven et al., 2000; Lyle & Smith, 2014; Macfarlan et al., 2012; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2004; Power & Ready, 2018).

The secondary line of investigation examined the potential effects participants' traits could have on concern for reputation scores (note that these refer to scores collected during the intake, not across the scheduled surveys). In line with predictions and prior studies (Combs & Penn, 2004; Lee & Jang, 2019), participants who had higher social interaction anxiety and social phobia were also more concerned for their reputation. Of the Big Five personality traits only agreeableness was significantly positively associated with concern for reputation. However, it is perhaps unsurprising that people who score high on a trait associated with motives to maintain positive interpersonal relations, are more concerned about their reputation (Graziano & Tobin, 2009). The other personality subscales and social perspective taking propensity were not significantly related. I also found no evidence to support the prediction that political conservatism would differentially affect levels of reputation concern felt across relationship types.

The present study used diary methods to extend the limits of experiments, and to study the common experience of reputation management in its natural setting, without restrictions on the behaviours, relationships or reputation domains sampled. Although this chosen method has the advantage of being able to test how reputation management varies across social relationships and contexts, it has the limitation of relying on the self-report of participants behaviours and feelings. Self-reports can introduce several biases and confounds. For instance, even though responses were anonymous, what participants chose to report—or not report—might have, ironically, been influenced by reputation management motives, resulting in selective reporting. Participants may have avoided describing situations perceived as embarrassing or morally questionable. This could be due to worries of there being a small chance their actions could become known, a desire to present themselves well even under anonymity, or a need to maintain a positive self-view. Participants could also have selectively reported more exciting or noteworthy events to avoid appearing uninteresting – or because they noticed and remembered such instances more –, which could skew data toward more salient or 'bigger' incidents. This

would leave more mundane instances underreported. (I should note, however, that many reported incidents did include mundane as well as embarrassing events, although I did not test for the frequency of cases.) Additionally, self-reports are constrained by participants' awareness of their motivations and behaviours. It is possible that they may not have fully understood or accurately articulated reputation management events. A lack of self-awareness can introduce bias, particularly if participants misinterpret or rationalise their behaviour in retrospect, leading to discrepancies between reported and actual behaviours, motivations or feelings. Therefore, it would be beneficial to consider these findings alongside other field methods that involve direct observations of reputation management, such as ethnographic methods and lab-in-the-field studies (Balliet et al., 2022). It would moreover be interesting to consider separate facets of reputation management to form a deeper understanding of reputation management in everyday life, as well as to turn more attention to specific behaviours – such as investigating differing reputational concerns during punishing behaviours or exploring whether we are good at predicting how our behaviours will affect our reputation.

The findings of this study are fairly comparable to previous reports from the laboratory and field. Even in everyday life, across various situational contexts and relationships, I found that people (i) care more about their reputation during reputation management events than during other social interactions, and (ii) that the social relationship aspects between people influence reputational concern, interestingly, perhaps more than social visibility cues. These findings highlight the need to further study the various aspects of social relationships, particularly the value that we place on a relationship, in relation to reputation concern.

# **Chapter 3**

## **The Reputation of Punishers in Daily Life**

### 3.1 Introduction

As previously established, punishment can bestow reputation benefits on punishers by increasing their likelihood of engaging in positive social interactions (Barclay, 2006; Batistoni et al., 2022; dos Santos et al., 2011, 2013; dos Santos & Wedekind, 2015; Jordan et al., 2016; Panchanathan & Boyd, 2004; Raihani & Bshary, 2015b, 2015c). Specifically, punishers can benefit by (i) signalling their formidability (dos Santos et al., 2013; Hilbe & Traulsen, 2012; Redhead et al., 2021) or (ii) signalling their cooperative intent (Barclay, 2006; Dhaliwal et al., 2020; Jordan et al., 2016; Nelissen et al., 2008; Raihani & Bshary, 2015c). Due to potential costs of retaliation (Nikiforakis, 2008; Rand et al., 2010), punishment can act as an honest signal of dominance or formidability (Gordon et al., 2014). Punishers can benefit from a tough reputation, as it discourages observers from refusing to cooperate in future (Clutton-Brock & Parker, 1995; dos Santos et al., 2011). Conversely, signalling cooperative intent via punishment (Jordan et al., 2016; Jordan & Rand, 2017; Przepiorka & Liebe, 2016) can lead to cooperative interactions with new social partners (e.g., Barclay, 2006; Dhaliwal et al., 2021; Jordan et al., 2016; Nelissen, 2008; Raihani & Bshary, 2015c). When punishment (i.e., the signal) is associated with cooperativeness (i.e., the punisher's hidden quality), observers can use this information to select the punisher as a partner (Batistoni et al., 2022; Jordan et al., 2016) or reward the punisher (Raihani & Bshary, 2015c). Although punishers can receive reputational benefits from punishing, it is a harmful act (Ohtsuki et al., 2009), and can therefore also be perceived as reflective of negative intentions (Fehr & Rockenbach, 2003; Nakamaru & Iwasa, 2006; Raihani et al., 2012; Raihani & Bshary, 2019). Besides encouraging cooperation, punishment can also function to increase payoffs or to improve punishers relative status (Bone et al., 2016; Deutchman et al., 2021; Herrmann et al., 2008; Raihani & Bshary, 2019). As the reputational consequences of punishment remain unclear (Maeda et al., 2022; Raihani & Bshary, 2015b), this study aimed to investigate several factors which may yield reputation benefits for punishers.

Raihani and Bshary (2015b) predicted that observers are likely to be sensitive to the punisher's motives and should only reward or select punishers as interaction partners when their punishing act signals that they were motivated by concerns for another's welfare, rather than by more selfish motives. Punishers are more often perceived as other-regarding versus self-serving when they themselves were not the victim of a violation and would also not benefit from a change in the target's behaviour – as generally occurs with third-party punishment (TPP). Consequently, and in line with a signalling account of punishment, third-party punishers are often perceived

as being more trustworthy (compared to second-party punishers, or individuals who do not punish) (Barclay, 2006; Batistoni et al., 2022; Jordan et al., 2016; Jordan & Rand, 2020; Kurzban et al., 2007; Nelissen, 2008; Raihani & Bshary, 2015b, 2015c) and may be rewarded (Raihani & Bshary, 2015c) or chosen as partners for interactions involving cooperation (Batistoni et al., 2022; Jordan et al., 2016; Nelissen, 2008), particularly when being chosen for a role that can provide benefits to the chooser (rather than for a role where the punisher would receive benefits; Horita, 2010). In contrast, when a punisher was the victim of a violation (i.e., second-party punishment (SPP)), observers are more likely to believe that the punishment stemmed from competitive rather than cooperative motives, leading to more negative perceptions of second-party punishers (Kriss et al., 2016; Raihani & Bshary, 2015b; Stüber, 2020). As SPP is often viewed as a more self-interested action whilst TPP is considered to be more prosocial (Martin et al., 2019), it has been proposed that TPP in particular is driven by selection pressures around the acquirement of a cooperative reputation (Kurzban et al., 2007). In line with this, SPP decisions are reportedly less affected by anonymity than TPP decisions (with TPP typically increasing more when observed), demonstrating a difference in observation and thereby reputation sensitivity (Bolton & Zwick, 1995; Fehr & Fischbacher, 2003; Kurzban et al., 2007).

Besides a punisher's perceived motivation and the type of punishment they employ (SPP, TPP, or a combination thereof), the conveyed information and its resulting reputational consequences are likely influenced by several additional context-specific factors: (i) how justified the punishment was, (ii) violation severity, (iii) whether the violation was intentional or accidental, (iv) the punisher's relationship to the violator, and (v) power dynamics. Individuals should discriminate between justified and unjustified punishment, so that only those who punish defectors are perceived positively and receive rewards (Panchanathan & Boyd, 2003). Indeed, observers do distinguish between justified and unjustified punishment, as punishers are only trusted more when the punishment was justified (Barclay, 2006) and unjustified punishment (aimed at fair individuals) is rewarded less often than justified punishment (Raihani & Bshary, 2015c). When evaluating how justified a punishment was, the proportionality of that punishment to the violation may be a core consideration (Bedau & Kelly, 2017). Generally, people believe that punitive responses should be proportional to the perceived severity of a violation (Carlsmith et al., 2002; Fehr & Gächter, 2000; Goodwin & Benforado, 2015; Hofmann et al., 2018). Even third-parties may be appraised as less interested in fairness restoration but rather as wanting to cause harm to the violator when punishment is excessive (Dhaliwal et al., 2021). In line with this, across societies punishment increases with the severity

of selfishness (Henrich et al., 2006) and observers only prefer third parties to help rather than punish when unfair actions are less severe (Li et al., 2021). Similarly, across 57 countries punitive responses were viewed increasingly favourably the more inappropriate a norm violation was perceived to be (Eriksson et al., 2021). Although non-action was overall perceived as the best response to norm violations, this was restricted to minor norm violations only – when violations were severe, inaction was no longer regarded as appropriate (Eriksson et al., 2021). In line with this, a study using vignettes determined that punishment and non-action were equally appropriate responses for minor norm violations, such as taking more than one's fair share of cake, but that for severe norm violations such as vandalism or speeding near children, only punitive responses were deemed to be appropriate (Eriksson et al., 2017). Generally, only individuals who carry certain roles, such as the police or teachers, are viewed as obligated to punish (Cushman, 2015; Eriksson et al., 2016; Guala, 2012), but when violations are severe adults and children believe that peers must punish as well (Marshall et al., 2020).

Additionally, Darley and Pittman (2003) suggested that both more severe and intentional violations result in punishment being considered a more normative action. Intentional violations tend to be viewed as more morally egregious and consequently receive harsher punishments compared to unintentional infractions (Cushman, 2008; Cushman et al., 2006; Greene et al., 2009; but see McNamara et al., 2020 for cross-cultural disparities in the significance of intent in moral evaluations). This may be because intentional versus accidental violations (as well as severe versus minor violations; Darley & Pittman, 2003; Reb et al., 2006) evoke greater moral outrage. Justified punishing behaviours in response to severe and intentional violations are therefore likely to receive the most favourable evaluations. Indeed, a vignette experiment conducted across eight small-scale societies and two Western societies revealed that intentional offenses were universally deemed more deserving of punishment compared to unintentional ones (Barrett et al., 2016). However, the degree to which intentionality influenced punishment judgments varied across cultures.

Details of the particular social relationships in question may also affect how a punisher is evaluated. For instance, observers believe it is more appropriate for those who are socially closer to the violator to confront them (Eriksson et al., 2017; Strimling & Eriksson, 2014). Looking at a combination of context factors, it has been found that punishers receive especially positive evaluations when a violator's friend punishes a severe violation, and especially negative evaluations when a stranger punishes a relatively minor violation (Eriksson et al., 2017).



The perception of punishment is additionally influenced by the status and power dynamics of the individuals involved. As people vary in resources and influence, power asymmetries can motivate some people to punish more than others (Hooper et al. 2010; Ruttan & Mulder Borgerhoff 1999). Observers tend to expect individuals with higher status to take on the role of punishers, as punishing is often more profitable and less risky for them (Gordon & Lea, 2016; Singh & Boomsma, 2015). Higher-status individuals can often inflict greater physical costs on targets (Sell et al., 2009) or limit access to resources (Keltner et al., 2003; Maner & Mead, 2010), and face reduced risks of retaliation (Gordon et al., 2014). Ethnographic studies also suggest that in societies with power asymmetries, individuals with higher political influence often adjudicate conflicts, leveraging their social coalitions to mitigate the risk of retaliation (von Rueden et al., 2012). High-status individuals who are expected to punish successfully are perceived as formidable and likeable (Gordon & Lea, 2016). However, as high-status individuals are expected to punish, failure to do so can jeopardize their status and reputation (Gordon & Lea, 2016; Price & Van Vugt, 2014). Moreover, whilst high-status punishers are more likely to be chosen as leaders than high-status non-punishers, whether an individual of lower status punishes or not has no effect on their likelihood of being selected as a leader (Gordon & Lea, 2016). It has therefore been suggested that only higher-status punishers can receive reputational benefits (Gordon & Lea, 2016). Conversely, when individuals are more equal, engaging in punishing behaviours may promote one's dominance status (Singh & Boomsma, 2015). Thus, the decision to punish is linked to social status and power dynamics, with higher-status individuals bearing the responsibility and expectation of maintaining social order through punitive actions.

It should again be noted that much of the research on the reputation of punishers has been conducted in the context of unjust economic interactions and we therefore know less about how the punishment of more personal violations is viewed (Dimitroff et al., 2022; Milfune et al., 2020; but see e.g. Hofmann et al., 2018; Molho et al., 2020). It is important to remember, that punishing behaviours and their evaluations in the laboratory differ from real-life social interactions as people's actions are not restricted to pre-determined options (Wu et al., 2022), and interactions involving others who are not strangers (which is predominantly the case in laboratory experiments) are frequent (Pedersen et al., 2020). The contextless nature of economic games can also result in defection being considered 'fair game', making the punishment of it less justified (Dhaliwal et al., 2021; Salcedo García, 2020). A lack of context, specifically a lack of direct information about a person's prior behaviour, is also less likely to occur in real life punishing situations compared to the laboratory. Conducting studies which

include context are important, as e.g. having direct information about a person's prior behaviour is occasionally treated as a caveat for punishers to receive reputation benefits (e.g., Dhaliwal et al., 2021; Jordan et al., 2017).

Therefore, to go beyond ensuring experimental control and investigating factors influencing a punishers reputation in isolation, I used a weekly diary methodology to gather information about people's experiences in daily life (Csikszentmihalyi & Larson, 2014). This method has high ecological validity as, rather than responding to restricted options in experiments or imagining response to a hypothetical scenario, it captures people's real responses to experiences and interactions in their lives (Molho et al., 2020; Uy et al., 2010). Additionally, rather than focusing on a single occurrence per person, this method enabled me to collect a variety of responses to several punishment incidents in everyday life. This study therefore investigated responses to punishing situations that vary in context, such as the degree of the norm violation, the relationship between punisher and violator, and differences in power dynamics, allowing my research to reflect the complexity of real-world environments.

I predicted that punishers' perceived motives would be strongly influenced by the type of punishment employed. Specifically, I expected that participants would attribute more cooperative motives to third-party punishers than to second-party punishers (Raihani & Bshary, 2015b) and that such punishers would be evaluated more positively. As punishing type and perceived motivation appear to be linked, I predicted that third-party punishers would receive more reputational benefits than second-party punishers or mixed punishers (i.e., those who punished when both they themselves and others were the victim of a violation, as occurs in public goods games). However, these reputation benefits may be specific to the cooperative reputation domain (i.e., third-party punishers will be viewed as more trustworthy, generous, cooperative, moral, and reliable, and participants will indicate a greater willingness to associate with or help them in future) as engaging in TPP may not inherently result in a more dominant, competent, or tough reputation than engaging in SPP. I also expected punishers to be evaluated more favourably when the punishment was perceived to be justified, the violation was severe, the violation was intentional rather than accidental, the punisher had a close relationship to the violator, and the punisher had more power.

## 3.2 Methods

### 3.2.1 Participants

I recruited 211 participants (158 females, 50 males, 3 preferred not to say) who were at least 18 years old ( $M = 38$  years,  $SD = 12.4$ , range = 18-70 years) and fluent in English via Prolific Academic to take part in six weekly surveys. Participants were predominantly White (89%) and British (82%), and many individuals were highly educated (58% had a bachelor's or higher degree). All participants provided informed consent before participating and the research was approved by the UCL Research Ethics Committee (Project ID: *ICN-PWB-NR-23-10-20A*). The median duration to complete each survey was 7 minutes. Participants were paid £0.70 per survey and could earn an additional bonus for completing five (£0.50) or all six surveys (£0.60). On average, participants received £3.43 ( $SD = 1.46$ ).

### 3.2.2 Weekly Surveys

A total of six surveys implemented in Qualtrics were made available once a week across a six-week period. These weekly surveys were made available on the same weekday for 24 hours, and asked the same questions about the reputation of punishers. I chose to schedule these surveys once a week based on the piloted frequency of which people witness or learn about punishing events in real life<sup>2</sup>. Leaving a 24-hour response window for each surveys gave participants the flexibility to respond at a time that suited them most, which I hoped would reduce attrition.

The first survey contained three main blocks: (1) an information and comprehension section in which participants received information about the study content and requirements, as well as a definition of punishment and related terms, a descriptive example, and a comprehension check; (2) a section collecting demographic information; and (3) questions about a punishment event participants had recently witnessed or learned about. If a participant reported that they did not witness or learn about a punishment event, they answered equivalent questions about a ‘non-

---

<sup>2</sup> Piloting for this survey involved testing it with a group of around 60 friends, family members, neighbours, and friends of friends or family (some of whom also participated in the pilot mentioned in the previous chapter). Participants were additionally asked to reflect on how frequently they observed or heard about punishing events in everyday life. This feedback informed decisions about the survey frequency to better align it with participants' real-world experiences of witnessing or learning about such events. The goal was to maximise response rates without overwhelming participants or causing them to struggle to recall examples, which could have led to poorer-quality data.

real life' punishing event instead (e.g., based on a punitive action they observed in a film or read in a novel). The subsequent surveys omitted the demographics section. See Appendix B1 for the exact survey content.

Participants completed between 1 and 6 surveys with a median response rate of 5 surveys ( $M = 4.42$ ,  $SD = 1.72$ ), resulting in  $k = 934$  responses. I subsequently excluded 29 of these responses from my analyses for failing to follow instructions: 10 responses were removed due to participants reporting on their own punishing behaviour, 14 for failing to describe a punishing event (for instance describing bullying situations instead), and 5 for other reasons, such as focusing on country-level sanctions. Due to contradictory information presented by participant, the 'punishment type' (i.e., SPP, TPP, or mixed) could not be determined for a further 45 responses, leaving  $k = 860$  viable responses. Participants reportedly witnessed or learned about a punishing event in 74% of these cases ( $k = 673$ ) and answered similar questions about punishing events from alternative 'non real-life' sources such as novels or films in the remaining cases ( $k = 232$ ).

### 3.2.3 Survey Content

The information section, which was repeated at the beginning of each survey, explained the concept of punishment, and highlighted the roles of the punisher, violator, and victim. Across several pages participants were told, that *"In our daily lives we punish people who violate social norms and expectations to varying degrees. Violations can be serious (e.g. stealing) or mundane (e.g. making an inconsiderate comment), and can be committed by someone familiar or a stranger. In this study, we consider punishment as a behaviour that inflicts a cost on (or removes benefits from) a person in response to a specific behaviour by that person. Punishment can take many forms, such as directly confronting a violator (either physically or verbally), sharing negative information about the violator to others (i.e., gossiping), demanding monetary compensation (i.e., material punishment), discipline (e.g., grounding a child), excluding the violator from social interactions or avoiding social contact with them. Within this survey we will refer to the person who engaged in the wrong behaviour as the 'violator'. We will refer to the person who was harmed or wronged by the violator as the 'victim'. The punisher is the one who punishes the violator for their wrong behaviour. The victim could be you yourself, the punisher, another person, or even the general community such as when a park is being littered. The punisher is not the violator. The violator is not the victim. The punisher can be the victim.*

*You can be the violator, the victim, or a bystander. You cannot be the punisher. Witnessing or learning about punishment acts include situations in which you yourself were punished, you watched someone else (e.g., a friend or a stranger) being punished in person or online, or you learned about a punishment act (e.g., because someone told or texted you about it, or you overheard gossip). Please focus on people rather than institution- or country-level punishment.”.*

Before commencing to the question section of the survey, participants had to select that they understood that (1) punishment occurs in response to a wrong behaviour, (2) punishment inflicts a cost on (or removes benefits from) the violator, (3) the punisher is not the violator, (4) the punisher can be the victim, (5) the violator cannot be the victim, and (6), they should not report their own punishing behaviour. If participants failed to do so they had to review the information again.

Subsequently, all participants were shown an example scenario, in which the correct relationships/ roles to be selected were explained. Specifically: *“In a scenario where ‘my friends’ child swore at her, so she punished him by cancelling pizza night’, the violation is swearing and the punishment is cancelling pizza night. The violator is the child. The punisher is the parent. The victim is the parent. The relationship between the punisher (parent) and the violator (child) is ‘family’. For the relationship between punisher and victim you would select ‘the punisher was the victim’, as in this case the parent was both the punisher AND the victim”.* Participants were then asked to recall a recent punishment act. If they could recall multiple punishing instances, they were asked to choose only one punishing act to report on throughout the survey. Participants were reminded to refrain from reporting their own punishing behaviours. Finally, they were asked whether they had witnessed or learned about a punishment act within the last week.

Whenever a participant indicated that they *had* witnessed or learned about a punishment act within the last week, they were asked to describe the punishing act. Subsequently, participants reported what form the punishment took (material punishment, avoidance, exclusion, withdrawal of benefits, verbal confrontation, physical confrontation, discipline, or other). Using two slider scales (with hidden values from 0 – 100 ranging from ‘*Not at all*’ to ‘*Very*’), participants then indicated how severe they perceived the violation that was punished to be and how justified they thought the punishment was. Following this, participants indicated whether they believed the violation was accidental or intentional. They also selected whether they themselves or someone else had been the violator. Similarly, they were asked whether they

themselves, the punisher, another person, or the general community/ environment had been the victim, selecting all responses that applied. Subsequently, I assessed what relationship the punisher and the violator, as well as the punisher and the victim had to each other (strangers, acquaintances, colleagues/classmates, competitors, friends, romantic partners, family, or other, with an additional option of 'the punisher was the victim' for the punisher-victim relationship question). I determined the punishing type (TPP, SPP or mixed punishment) based on who participants selected as the victim(s) of a violation. When the punisher was not selected as a victim, I coded it as TPP. When only the punisher was selected as the victim, it was classed as SPP. When both the punisher and another person or the general community/ environment were the victims, punishment was classed as mixed. Further looking into relationship contexts, participants also indicated whether there was any power asymmetry between the punisher and the violator on a sliding scale (from '*The punisher had more power*' to '*The violator had more power*' with the midpoint labelled as '*Equal*').

I then assessed the motivation participants attributed to punishers ('*Do you think that the punisher's motivation for punishing was more altruistic or more self-serving?*'), with a sliding scale from '*Altruistic*' to '*Self-Serving*'. Again using slider scales, participants then provided their perception of the punisher, specifically how likeable, trustworthy, generous, cooperative, dominant, moral, competent, reliable, and tough they believed the punisher to be. As an overall perception of the punishing event, participants indicated whether they thought that punishing reflected on the punisher more negatively or more positively (from '*Definitely negatively*' to '*Definitely positively*' with '*Neither*' as midpoint). Finally, I measured behavioural intentions by asking whether, if the opportunity were to present itself in future, the participant would want to help the punisher (with a sliding scale from '*not want to help the punisher*' to '*want to help the punisher*'), and whether the punishing behaviour decreased or increased their willingness to associate (e.g., be friends) with the punisher (from '*Definitely decrease*' to '*Definitely increase*' with '*Neither*' as midpoint label).

Whenever participants could not report on a real-life punishing event, they answered equivalent questions about a 'non-real life' punishing event they encountered instead. This alternative question block additionally asked about the source of the 'non real-life' punishing event the participant was reporting on (e.g., a comic or a fictional TV show). The alternative question block was included to prevent participants from simply selecting 'no' to purposefully cut the survey short.

### 3.2.4 Analysis Plan

Only responses on real-life punishing events were included in the analyses (see Appendix B2 for results from analyses that include ‘non real-life’ punishing instances from e.g., novels or films as well). Analyses were conducted in R (R Core Team, 2020) using the packages ‘tidyr’ (Wickham & Girlich, 2022), ‘lme4’ (Bates et al., 2015), ‘lmerTest’ (Kuznetsova, 2017), ‘MuMIn’ (Barton, 2022), ‘emmeans’ (Lenth, 2022), and ‘sjPlot’ (Lüdtke, 2021). All reported coefficients are unstandardised.

To examine whether punishers were perceived more positively or negatively, ratings for all twelve response variables were centred around the scale midpoint of 50. These adjusted ratings were then employed as the response variables in twelve constant (intercept-only) linear mixed models with participant ID as random effect. As twelve models (one for each evaluation item) were fit, I applied a Bonferroni correction to adjust for multiple comparisons, leading to an adjusted significance level of  $\alpha = .0042$ .

To explore the factors that influence how a punisher is evaluated, I used the R package ‘MuMIn’ to analyse the data using multi-model selection with model averaging (Grueber et al., 2011). Model selection involves identifying the best-fitting model from a set of candidates, while model averaging incorporates multiple models simultaneously to infer parameters. This approach can offer some advantages over more traditional methods. Burnham et al. (2011) for instance caution against relying solely on a global model with all predictors. A global model can lead to imprecise parameter estimates, becoming unstable and potentially uninformative as the number of parameters increases. This instability raises the risk of detecting spurious effects and increases uncertainty in new parameter estimates (Burnham et al., 2011).

The goal of model selection is to balance model fit and complexity, minimising the risk of overfitting or oversimplifying the data (Harrison et al., 2018). Typically, this involves comparing a candidate model containing a term of interest to a ‘null’ model without that term, using null-hypothesis significance testing (NHST; Nickerson, 2000). A common method is stepwise deletion, which sequentially removes terms from a global model to derive a ‘minimal adequate model.’ Only significant predictors are retained, as determined through NHST (Crawley, 2013). However, stepwise deletion has been criticised for overlooking model uncertainty and potentially overestimating the effect sizes of significant predictors (Burnham et al., 2011; Forstmeier & Schielzeth, 2011; Whittingham et al., 2006).

An Information-Theoretic (IT) framework addresses these limitations by ranking and weighting multiple competing models, providing a quantitative measure of their relative support (Grueber et al., 2011; Harrison et al., 2018). This approach, often based on metrics like the Akaike Information Criterion (AIC), accounts for model selection uncertainty by recognising that several models may fit the data similarly well (Burnham & Anderson, 2002; Burnham et al., 2011). It is particularly useful when models have comparable complexity, such as identical numbers of predictors, where traditional NHST is inadequate due to its reliance on comparisons with simpler, nested models (Harrison et al., 2018). Especially when predictors exhibit complex, context-dependent relationships, focusing solely on one model risks overlooking important insights. Multi-model inference mitigates this risk by ranking models based on their relative support and deriving more stable parameter estimates through averaging across multiple models (Grueber et al., 2011; Harrison et al., 2018; Johnson & Omland, 2004; Symonds & Moussalli, 2011).

Simulation studies have shown that although single 'best' model selection can yield similar parameter estimates, model-averaged results are generally more robust and less prone to dismissing weak but potentially important effects (Richards, 2005; Richards et al., 2011). As outlined above, overreliance on a single 'best' model disregards model selection uncertainty, potentially omitting valuable insights found in other competing models (Burnham et al., 2011; Chatfield, 1995; Draper, 1995; Yuan & Yang, 2005). By incorporating model uncertainty into parameter estimates, model averaging offers more reliable estimates and robust confidence intervals than single-model inference (Fletcher & Fletcher, 2018). Consequently, model averaging can reduce variability in model selection and mitigate modelling biases, providing a more stable basis for inference (Peng & Yang, 2022).

Practically, the process generally begins with constructing a global model that includes all relevant predictors, ensuring they are theoretically justified and do not reflect implausible or mutually exclusive relationships (Grueber et al., 2011; Dochtermann & Jenkins, 2010). From this global model, a "model set" is generated, typically consisting of all possible submodels, including an intercept-only model (Grueber et al., 2011). These submodels are then compared using an information criterion, such as the AIC, to rank their relative support. Commonly, a "top model set" is identified, often comprising models with AIC differences of less than 2 relative to the best model, as these are considered to have comparable support in the data (Burnham & Anderson, 2002). While the all-subset approach, involving all possible combinations of predictors, has been critiqued as "data-dredging" (Burnham & Anderson,



2002), it remains appropriate when there is a reasonable basis to assume that all predictors could plausibly influence the outcome of interest (Dochtermann & Jenkins, 2011).

Additionally, collinearity among predictors poses challenges for model interpretation and parameter estimation. When predictors are highly correlated, they share variance in the response variable, potentially leading to model convergence issues, inflated standard errors, and incorrect effect interpretations (Graham, 2003; Quinn & Keough, 2002; Zuur et al., 2010). In model selection, high collinearity complicates identifying the ‘best’ model, as different models with correlated predictors may fit similarly well. To address these issues, it is important to assess multicollinearity before fitting models (Zuur & Ieno, 2016) and performing model averaging (Cade, 2015; Lukacs et al., 2010). According to Dormann et al. (2013), correlation coefficients exceeding 0.7 indicate that the variables may not be suitable for inclusion in the same model due to redundancy. In the present study, absolute correlation values ranged from 0.05 (between violation severity and power dynamics) to 0.53 (between perceived justness of punishment and selfishness in punishing motives). I also ensured that predictors were on common scales, reducing potential issues with model-averaged coefficients (Cade, 2015; Zuur & Ieno, 2016). Consistent scaling facilitates meaningful interpretation of model-averaged estimates while minimising biases from collinearity (Freckleton, 2010). Since no correlations exceeded the 0.7 threshold, no predictors were dropped, but it is important to acknowledge that high correlations can still complicate the interpretation of results.

Following common procedure, for each reputational outcome I first specified a global model, incorporating all relevant predictors (i.e., perceived motivation, punishing type, justness of the punishment, violation intentionality, violation severity, power dynamics, and the punisher-violator relationship) as fixed effects. Participant ID was set as a random effect in all models. I then performed automated model selection by generating all possible subsets of the global model, including the null model, to identify a set of top models. I further narrowed this set of top models to include only those models that were within 2 AICc units of the ‘best’ model. Lastly, I averaged across these refined top model sets to obtain parameter estimates to reduce uncertainty about which model actually is the ‘best’ model (Burnham & Anderson, 2004). The same method was used to explore which factors influenced a punisher’s perceived motivation. As thirteen models (one for each evaluation item, and an additional model for perceived motivation) were fit, I applied a Bonferroni correction to adjust for multiple comparisons. The significance level ( $\alpha = .05$ ) was divided by 13 to account for the number of comparisons being made, leading to an adjusted significance level of  $\alpha = .0038$ .

To address potential criticisms of model averaging, particularly regarding challenges in interpreting effects that may vary across models (Cade, 2015), I included the results from the global models—without model selection or averaging—in Appendix B3. Notably, only two coefficients showed changes in significance, demonstrating that the findings remained consistent and robust across different analytical approaches.

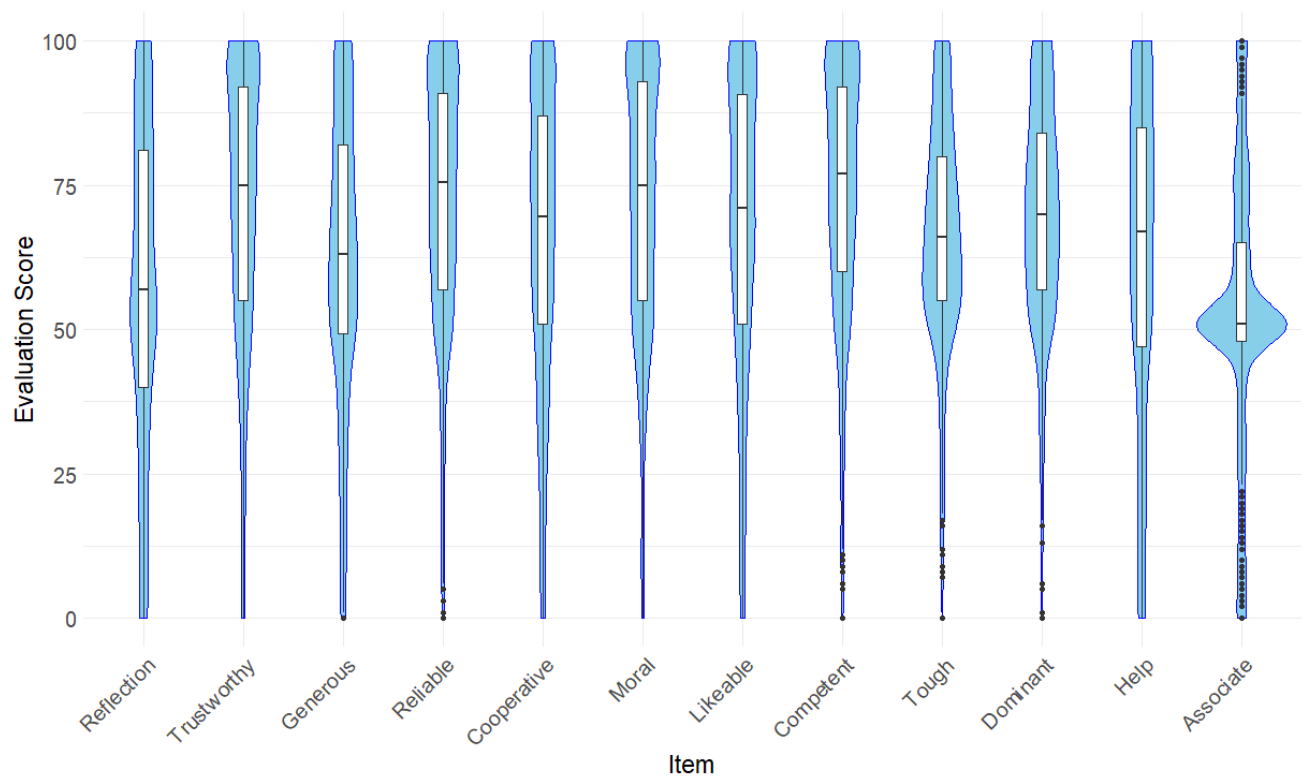
Lastly, to assess whether punishing provided distinct formidable or prosocial reputation benefits depending on the punisher's relationship to the target, I conducted a mixed-effects Analysis of Variance (ANOVA) using Satterthwaite's method to predict observer judgments as a function of the relationship between punisher and violator, reputation type (formidable or prosocial), and their interaction. Subsequently, I performed pairwise comparisons on the estimated marginal means ("emmeans" package; Lenth, 2023) for the interaction between relationship and reputation. Note that for this analysis, 'prosocial reputation' was the average row score of trustworthiness, generosity, likeability, cooperativeness, morality and reliability, while 'formidable reputation' was the average row score of competence, dominance and toughness.

### 3.3 Results

On average, punishers were rated relatively positively for their punishing behaviours (Figure 3.1). Constant (intercept-only) linear mixed models showed that punishers were perceived as competent ( $b = 23.16$ , 95% CI [20.91, 25.39],  $t(174.20) = 20.35$ ,  $p < .0001$ ), reliable ( $b = 21.30$ , 95% CI [19.04, 23.54],  $t(175.96) = 18.64$ ,  $p < .0001$ ), moral ( $b = 21.17$ , 95% CI [18.77, 23.55],  $t(166.36) = 17.43$ ,  $p < .0001$ ), trustworthy ( $b = 20.67$ , 95% CI [18.26, 23.05],  $t(176.47) = 16.96$ ,  $p < .0001$ ), dominant ( $b = 19.50$ , 95% CI [17.72, 21.27],  $t(160.25) = 21.64$ ,  $p < .0001$ ), likeable ( $b = 17.57$ , 95% CI [15.02, 20.12],  $t(181.63) = 13.55$ ,  $p < .0001$ ), tough ( $b = 16.20$ , 95% CI [14.25, 18.14],  $t(168.61) = 16.39$ ,  $p < .0001$ ), cooperative ( $b = 15.08$ , 95% CI [12.52, 17.62],  $t(174.65) = 11.66$ ,  $p < .0001$ ), and generous ( $b = 11.67$ , 95% CI [9.24, 14.10],  $t(179.24) = 9.43$ ,  $p < .0001$ ). Observers also indicated that they were significantly more likely to help rather than not help punishers in future ( $b = 12.42$ , 95% CI [9.71, 15.11],  $t(183.94) = 9.03$ ,  $p < .0001$ ). Lastly, although observers indicated that punishing reflected more positively than negatively on the punisher overall ( $b = 7.62$ , 95% CI [4.92, 10.32],  $t(184.40) = 5.55$ ,  $p < .0001$ ), punishment did not significantly increase willingness to associate with the punisher in future ( $b = 2.99$ , 95% CI [0.82, 5.14],  $t(169.78) = 2.72$ ,  $p = .007$ ).

#### Figure 3.1

*Average Perceptions of Punishers Across Various Dimensions*



*Note.* Violin plots illustrate the distribution of observer perceptions of punishers across all twelve evaluation items. Boxplots provide information on median (horizontal line), interquartile range (IQR) (box), and outliers (points). Significance levels are indicated above violin plots with \*\*\* denoting  $p < .0001$  and "n.s"  $p > .004$ . Higher scores indicate higher evaluations of punishers on the relevant trait. 'Reflection' refers to observers' overall perception of punishers, specifically whether they thought punishing reflected on the punisher more negatively (lower values) or more positively (higher values). 'Help' and 'Associate' represent observers' willingness to help or associate with the punisher in future, with higher values indicating higher willingness.

Following the analysis of average punisher perceptions, correlations among response variables were examined to further understand the relationships between evaluation items (see Table 3.1). The strongest correlations were between reliability and competence ( $r = 0.88$ ), as well as trustworthiness and likeability ( $r = 0.84$ ). Additionally, cooperativeness was strongly correlated with both likeability and trustworthiness (both  $r = 0.81$ ). In contrast, the weakest correlations involved dominance, showing minimal associations with reliability ( $r = 0.02$ ), competence ( $r = 0.03$ ), and morality ( $r = -0.03$ ).

**Table 3.1***Correlation Matrix of Response Variables Relating to Punisher Perceptions*

|             | Likeable | Trustworthy | Generous | Cooperative | Dominant | Moral | Competent | Reliable | Strong | Overall | Helping | Associate |
|-------------|----------|-------------|----------|-------------|----------|-------|-----------|----------|--------|---------|---------|-----------|
| Likeable    | --       |             |          |             |          |       |           |          |        |         |         |           |
| Trustworthy | 0.84     | --          |          |             |          |       |           |          |        |         |         |           |
| Generous    | 0.77     | 0.73        | --       |             |          |       |           |          |        |         |         |           |
| Cooperative | 0.81     | 0.81        | 0.76     | --          |          |       |           |          |        |         |         |           |
| Dominant    | -0.16    | -0.07       | -0.19    | -0.13       | --       |       |           |          |        |         |         |           |
| Moral       | 0.77     | 0.8         | 0.68     | 0.77        | -0.03    | --    |           |          |        |         |         |           |
| Competent   | 0.71     | 0.78        | 0.63     | 0.73        | 0.03     | 0.78  | --        |          |        |         |         |           |
| Reliable    | 0.70     | 0.80        | 0.64     | 0.76        | 0.02     | 0.78  | 0.88      | --       |        |         |         |           |
| Strong      | 0.23     | 0.30        | 0.15     | 0.22        | 0.41     | 0.3   | 0.4       | 0.39     | --     |         |         |           |
| Overall     | 0.66     | 0.64        | 0.55     | 0.65        | -0.07    | 0.66  | 0.56      | 0.58     | 0.24   | --      |         |           |
| Helping     | 0.72     | 0.68        | 0.59     | 0.64        | -0.15    | 0.65  | 0.58      | 0.6      | 0.21   | 0.76    | --      |           |
| Associate   | 0.70     | 0.64        | 0.58     | 0.64        | -0.11    | 0.62  | 0.54      | 0.56     | 0.24   | 0.73    | 0.74    | --        |

*Note.* This table presents the Pearson correlation coefficients among response variables related to the perception of punishers. The variables include likeability, trustworthiness, generosity, cooperativeness, dominance, morality, competence, reliability, strength, the overall reflection of punishment on the target (with higher values indicating a more positive perception), willingness to help the punisher, and willingness to associate with the punisher.

Additionally, examining the frequencies of categorical variables offers valuable context. In terms of punishment types, there were 360 instances of third-party punishment, 201 instances of second-party punishment, and 77 instances involving a mix of second- and third-party punishment. The most frequently mentioned violator-punisher relationships included family ( $k = 176$ ), strangers ( $k = 147$ ), colleagues ( $k = 108$ ), others ( $k = 82$ ), acquaintances ( $k = 48$ ), friends ( $k = 44$ ), romantic partners ( $k = 28$ ), and lastly competitors ( $k = 5$ ). Additionally, participants reportedly believed that the violation being punished had been intentional in 464 instances versus accidental in 174 instances.

The subsequent sections will detail the impact of all predictors on different dimensions of a punisher's reputation, starting with participants' overall judgement of the punishment (i.e., whether, overall, they perceive punishment as reflecting negatively or positively on the punisher). This will be followed by analyses of punishers' prosocial reputations, formidability reputations, and participant's behavioural intentions towards punishers. For a comprehensive overview of all factors influencing a punisher's reputation in each final model, please refer to Table 3.2.

### 3.3.1 Overall Reflection of Punishment on the Punisher

Punishers were perceived significantly more positively when the punishment was justified ( $b = 0.59$ , 95% CI [0.53, 0.66],  $p < .0001$ ) and when punishment was thought to be driven by more altruistic than selfish motives ( $b = -0.20$ , 95% CI [-0.25, -0.14],  $p < .0001$ ). Violation intentionality, violation severity, power dynamics, punishing type, and the relationship between violator and punisher did not significantly influence how positively punishing reflected on the punisher (all  $p > .05$ ).

### 3.3.2 The Influence of Contextual Factors on Punishers' Prosocial Reputations

Punishers were perceived as significantly more likable ( $b = 0.50$ , 95% CI [0.44, 0.56],  $p < .0001$ ), cooperative ( $b = 0.48$ , 95% CI [0.42, 0.54],  $p < .0001$ ), trustworthy ( $b = 0.46$ , 95% CI [0.40, 0.52],  $p < .0001$ ), moral ( $b = 0.44$ , 95% CI [0.39, 0.50],  $p < .0001$ ), generous ( $b = 0.40$ , 95% CI [0.33, 0.47],  $p < .0001$ ), and reliable ( $b = 0.38$ , 95% CI [0.32, 0.44],  $p < .0001$ ) when the punishment was thought to be justified.

They were also regarded as more cooperative ( $b = -0.20$ , 95% CI [-0.26, -0.15],  $p < .0001$ ), trustworthy ( $b = -0.20$ , 95% CI [-0.26, -0.15],  $p < .0001$ ), likeable ( $b = -0.18$ , 95% CI [-0.24, -0.12],  $p < .0001$ ), generous ( $b = -0.17$ , 95% CI [-0.24, -0.11],  $p < .0001$ ), moral ( $b = -0.14$ , 95% CI [-0.19, -0.09],  $p < .0001$ ) and reliable ( $b = -0.13$ , 95% CI [-0.18, -0.07],  $p < .0001$ ) when punishers were attributed more altruistic motives.

A punisher's prosocial reputation was further influenced by the relationship between punisher and violator. Punishers were evaluated as significantly more generous ( $b = 16.17$ , 95% CI [11.66, 20.69],  $p < .0001$ ), likeable ( $b = 12.04$ , 95% CI [7.91, 16.18],  $p < .0001$ ), cooperative ( $b = 9.99$ , 95% CI [5.73, 14.25],  $p < .0001$ ), trustworthy ( $b = 9.54$ , 95% CI [5.46, 13.61],  $p < .0001$ ), moral ( $b = 8.83$ , 95% CI [5.11, 12.55],  $p < .0001$ ), and reliable ( $b = 7.84$ , 95% CI [3.66, 12.03],  $p < .0001$ ) when they punished a family member rather than a stranger (and also significantly more generous when a romantic partner ( $b = 19.60$ , 95% CI [11.35, 27.85],  $p < .0001$ ), friend ( $b = 14.23$ , 95% CI [7.64, 20.82],  $p < .0001$ ), or acquaintance ( $b = 9.95$ , 95% CI [3.57, 16.34],  $p = .002$ ) rather than a stranger was punished, and significantly more likeable when a romantic partner ( $b = 12.73$ , 95% CI [5.15, 20.31],  $p = .001$ ) or friend ( $b = 10.10$ , 95% CI [3.91, 16.29],  $p = .001$ ) as opposed to a stranger was punished).

**Table 3.2**

*The association between context variables and punishers' reputations*

|                                                 | Punishment<br>Reflection            | Trust-<br>worthiness                | Generosity                          | Reliability                         | Cooperative<br>-ness                | Morality                            | Likeability                         | Competence                          | Toughness                           | Dominance                           | Willingness<br>to help              | Willingness<br>to associate         |
|-------------------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| <i>Predictors</i>                               | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    |
| (Intercept)                                     | <b>21.06</b> ***<br>(14.76 – 27.36) | <b>40.28</b> ***<br>(33.46 – 47.10) | <b>27.02</b> ***<br>(19.44 – 34.60) | <b>46.06</b> ***<br>(38.34 – 53.77) | <b>32.35</b> ***<br>(25.04 – 39.66) | <b>41.02</b> ***<br>(34.63 – 47.41) | <b>30.69</b> ***<br>(23.66 – 37.73) | <b>52.05</b> ***<br>(45.80 – 58.30) | <b>65.09</b> ***<br>(56.15 – 74.04) | <b>81.76</b> ***<br>(75.22 – 88.30) | <b>16.45</b> ***<br>(8.43 – 24.48)  | <b>23.04</b> ***<br>(16.91 – 29.17) |
| Intentional<br>Violation                        | 3.14<br>(-0.19 – 6.48)              |                                     | 4.81<br>(1.10 – 8.52)               | 0.73<br>(-2.64 – 4.10)              | 3.82<br>(0.31 – 7.33)               | 1.95<br>(-1.12 – 5.02)              | 4.39<br>(0.92 – 7.86)               | 1.45<br>(-1.91 – 4.82)              |                                     |                                     | 3.86<br>(0.16 – 7.56)               | 4.33<br>(1.04 – 7.62)               |
| Justified<br>Punishment                         | <b>0.59</b> ***<br>(0.53 – 0.66)    | <b>0.46</b> ***<br>(0.40 – 0.52)    | <b>0.40</b> ***<br>(0.33 – 0.47)    | <b>0.38</b> ***<br>(0.32 – 0.44)    | <b>0.48</b> ***<br>(0.42 – 0.54)    | <b>0.44</b> ***<br>(0.39 – 0.50)    | <b>0.50</b> ***<br>(0.44 – 0.56)    | <b>0.36</b> ***<br>(0.30 – 0.42)    | 0.07<br>(0.00 – 0.13)               | -0.08<br>(-0.14 – -0.02)            | <b>0.59</b> ***<br>(0.52 – 0.65)    | <b>0.42</b> ***<br>(0.36 – 0.48)    |
| Punishing<br>Motive                             | <b>-0.20</b> ***<br>(-0.25 – -0.14) | <b>-0.20</b> ***<br>(-0.26 – -0.15) | <b>-0.17</b> ***<br>(-0.24 – -0.11) | <b>-0.13</b> ***<br>(-0.18 – -0.07) | <b>-0.20</b> ***<br>(-0.26 – -0.15) | <b>-0.14</b> ***<br>(-0.19 – -0.09) | <b>-0.18</b> ***<br>(-0.24 – -0.12) | <b>-0.14</b> ***<br>(-0.20 – -0.09) | -0.06<br>(-0.12 – -0.00)            | 0.02<br>(-0.03 – 0.08)              | <b>-0.16</b> ***<br>(-0.22 – -0.10) | <b>-0.13</b> ***<br>(-0.18 – -0.07) |
| Violation<br>Severity                           | 0.02<br>(-0.03 – 0.08)              | 0.01<br>(-0.04 – 0.07)              | 0.03<br>(-0.04 – 0.09)              |                                     | -0.02<br>(-0.08 – 0.04)             | 0.02<br>(-0.04 – 0.07)              | -0.02<br>(-0.08 – 0.04)             |                                     | <b>0.12</b> ***<br>(0.06 – 0.18)    | 0.05<br>(-0.01 – 0.11)              | 0.01<br>(-0.05 – 0.08)              | 0.02<br>(-0.03 – 0.07)              |
| Power<br>Dynamics                               | 0.02<br>(-0.04 – 0.08)              | -0.04<br>(-0.10 – 0.02)             | 0.03<br>(-0.03 – 0.10)              | -0.05<br>(-0.12 – 0.01)             | -0.04<br>(-0.10 – 0.03)             | 0.01<br>(-0.04 – 0.07)              |                                     | -0.05<br>(-0.10 – 0.01)             | <b>-0.12</b> **<br>(-0.18 – -0.06)  | <b>-0.19</b> ***<br>(-0.26 – -0.13) | 0.07<br>(-0.00 – 0.13)              | 0.06<br>(0.01 – 0.12)               |
| Mixed<br>Punishment                             |                                     | 5.29<br>(0.96 – 9.62)               | 3.03<br>(-1.78 – 7.84)              | 4.25<br>(-0.25 – 8.76)              |                                     |                                     |                                     | 2.26<br>(-2.09 – 6.61)              |                                     |                                     |                                     | 3.34<br>(-0.87 – 7.54)              |
| SPP                                             |                                     | 2.92<br>(-0.41 – 6.25)              | 3.40<br>(-0.33 – 7.12)              | 1.87<br>(-1.57 – 5.30)              |                                     |                                     |                                     | 4.18<br>(0.95 – 7.41)               |                                     |                                     |                                     | -1.69<br>(-4.82 – 1.44)             |
| Violator<br>Relationship<br>[Acquaintanc<br>es] |                                     | 4.42<br>(-1.34 – 10.19)             | <b>9.95</b> *<br>(3.57 – 16.34)     | 1.46<br>(-4.44 – 7.35)              | 3.82<br>(-2.31 – 9.95)              | -1.54<br>(-6.93 – 3.85)             | 4.93<br>(-1.12 – 10.99)             |                                     | -0.39<br>(-6.37 – 5.60)             | 1.64<br>(-4.56 – 7.84)              | 5.80<br>(-0.66 – 12.25)             |                                     |

|                                                         | Punishment<br>Reflection                      | Trust-<br>worthiness                            | Generosity                                   | Reliability                                   | Cooperative<br>-ness                          | Morality                                       | Likeability      | Competence       | Toughness                                      | Dominance                                      | Willingness<br>to help                         | Willingness<br>to associate |
|---------------------------------------------------------|-----------------------------------------------|-------------------------------------------------|----------------------------------------------|-----------------------------------------------|-----------------------------------------------|------------------------------------------------|------------------|------------------|------------------------------------------------|------------------------------------------------|------------------------------------------------|-----------------------------|
| <i>Predictors</i>                                       | <i>Estimates</i>                              | <i>Estimates</i>                                | <i>Estimates</i>                             | <i>Estimates</i>                              | <i>Estimates</i>                              | <i>Estimates</i>                               | <i>Estimates</i> | <i>Estimates</i> | <i>Estimates</i>                               | <i>Estimates</i>                               | <i>Estimates</i>                               | <i>Estimates</i>            |
| Violator<br>Relationship<br>[Colleagues/<br>Classmates] | 3.74<br>(-<br>0.73 – 8.20)                    | 6.26<br>(1.32 – 11.20<br>)                      | 6.63<br>(2.05 – 11.21<br>)                   | 6.06<br>(1.31 – 10.80<br>)                    | 1.39<br>(-<br>2.78 – 5.57)                    | 3.36<br>(-<br>1.29 – 8.01)                     |                  |                  | -1.96<br>(-<br>6.59 – 2.66)                    | -3.07<br>(-<br>7.83 – 1.69)                    | 2.88<br>(-<br>2.14 – 7.90)                     |                             |
| Violator<br>Relationship<br>[Competitors]               | -6.43<br>(-<br>21.74 – 8.87)                  | 4.05<br>(-<br>12.84 – 20.9<br>4)                | -0.58<br>(-<br>16.13 – 14.9<br>7)            | -1.32<br>(-<br>17.52 – 14.8<br>9)             | -6.59<br>(-<br>20.72 – 7.54)                  | -11.75<br>(-<br>27.80 – 4.31)                  |                  |                  | 4.40<br>(-<br>11.45 – 20.2<br>5)               | 6.56<br>(-<br>10.09 – 23.2<br>0)               | 12.60<br>(-<br>4.41 – 29.61<br>)               |                             |
| Violator<br>Relationship<br>[Family]                    | <b>9.54 ***</b><br><b>(5.46 – 13.61<br/>)</b> | <b>16.17 ***</b><br><b>(11.66 – 20.6<br/>9)</b> | <b>7.84 **</b><br><b>(3.66 – 12.03<br/>)</b> | <b>9.99 ***</b><br><b>(5.73 – 14.25<br/>)</b> | <b>8.83 ***</b><br><b>(5.11 – 12.55<br/>)</b> | <b>12.04 ***</b><br><b>(7.91 – 16.18<br/>)</b> |                  |                  | <b>-6.30 *</b><br><b>(-10.54 – -<br/>2.07)</b> | <b>-6.69 *</b><br><b>(-11.06 – -<br/>2.32)</b> | <b>10.86 ***</b><br><b>(6.24 – 15.4<br/>8)</b> |                             |
| Violator<br>Relationship<br>[Friends]                   | 5.41<br>(-<br>0.53 – 11.36)                   | <b>14.23 ***</b><br><b>(7.64 – 20.82<br/>)</b>  | 3.96<br>(-<br>2.11 – 10.03)                  | 5.04<br>(-<br>1.24 – 11.31)                   | 5.75<br>(0.25 – 11.25<br>)                    | <b>10.10 *</b><br><b>(3.91 – 16.29<br/>)</b>   |                  |                  | -4.27<br>(-<br>10.38 – 1.84)                   | -5.32<br>(-<br>11.68 – 1.04)                   | <b>10.77 *</b><br><b>(4.17 – 17.3<br/>8)</b>   |                             |
| Violator<br>Relationship<br>[Other]                     | 4.94<br>(0.04 – 9.85)                         | 2.21<br>(-<br>3.19 – 7.61)                      | 2.99<br>(-<br>2.08 – 8.05)                   | -2.02<br>(-<br>7.23 – 3.19)                   | 3.57<br>(-<br>0.98 – 8.12)                    | 1.88<br>(-<br>3.16 – 6.92)                     |                  |                  | 2.24<br>(-<br>2.89 – 7.37)                     | 2.92<br>(-<br>2.36 – 8.20)                     | 7.09<br>(1.48 – 12.7<br>0)                     |                             |
| Violator<br>Relationship<br>[Romantic<br>Partners]      | 7.50<br>(0.09 – 14.91<br>)                    | <b>19.60 ***</b><br><b>(11.35 – 27.8<br/>5)</b> | 2.74<br>(-<br>4.76 – 10.23)                  | 9.69<br>(1.99 – 17.39<br>)                    | 4.49<br>(-<br>2.30 – 11.28)                   | <b>12.73 **</b><br><b>(5.15 – 20.31<br/>)</b>  |                  |                  | -4.30<br>(-<br>11.84 – 3.24)                   | -10.08<br>(-17.78 – -<br>2.39)                 | <b>13.81 **</b><br><b>(5.69 – 21.9<br/>3)</b>  |                             |
| N                                                       | 190 ID                                        | 190 ID                                          | 190 ID                                       | 190 ID                                        | 190 ID                                        | 190 ID                                         | 190 ID           | 190 ID           | 190 ID                                         | 190 ID                                         | 190 ID                                         | 190 ID                      |
| Observations                                            | 638                                           | 638                                             | 638                                          | 638                                           | 638                                           | 638                                            | 638              | 638              | 638                                            | 638                                            | 637                                            | 637                         |

\*  $p < 0.0038$  \*\*  $p < 0.001$  \*\*\*  $p < 0.0001$

*Note.* Missing cells indicate that the specific factor was not included in the final top model. Estimates reported are conditional model averages, which average over the models where the parameters appear. The reference level for Violator Relationship is “Strangers”, “TPP” for Punishing Type, and “Accidental” for Violation Intentionality. "Punishment Reflection" assessed whether observers viewed punishment as reflecting positively or negatively on the punisher overall. For “Punishing Motive” higher estimates signify perceptions of more self-serving motives, while



lower estimates signify perceptions of more altruistic punishing motives. Lower “Power Dynamics” scores indicate the punisher held more power, while higher scores indicate that the violator held more power. Significant results are bolded. Estimates are followed by 95% confidence intervals (CIs) in brackets.

A punisher’s perceived motivation was influenced by the punishing type they engaged in, as well as by how justified the punishment was judged to be. Punishers were judged as being significantly more driven by self-serving than altruistic motives when they engaged in SPP compared to TPP ( $b = 15.76$ , 95% CI [11.54, 19.97],  $p < .0001$ ), and a punisher’s motives were also perceived as significantly less self-serving when the punishment was viewed as justified ( $b = -0.51$ , 95% CI [-0.58, -0.44],  $p < .0001$ ) (see Table 3.3 for the full model summary). Refer to Appendix B4 for results from analyses in which interactions between motive and punishing type are included in the twelve reputation models.

**Table 3.3**

*Context variables associated with perceptions of punishing motives*

| <i>Predictors</i>     | <b>Perceptions of Punishing Motive</b> |               |                   |
|-----------------------|----------------------------------------|---------------|-------------------|
|                       | <i>b</i>                               | <i>95% CI</i> | <i>p</i>          |
| (Intercept)           | 72.92                                  | 66.34 – 79.50 | <b>&lt;0.0001</b> |
| Justified Punishment  | -0.51                                  | -0.58 – -0.44 | <b>&lt;0.0001</b> |
| Power Dynamics        | 0.07                                   | -0.00 – 0.15  | 0.061             |
| Mixed Punishment      | 5.45                                   | -0.45 – 11.35 | 0.070             |
| SPP                   | 15.74                                  | 11.52 – 19.95 | <b>&lt;0.0001</b> |
| Violation Severity    | -0.03                                  | -0.11 – 0.04  | 0.411             |
| Intentional Violation | -1.36                                  | -5.94 – 3.21  | 0.559             |
| N <sub>ID</sub>       | 190                                    |               |                   |
| Observations          | 638                                    |               |                   |

*Note.* The relationship between the punisher and violator was not included in the final top model and is therefore absent from the table. Estimates reported are conditional model averages, which average over the models where the parameters appear. Significant results are bolded. Higher estimates signify perceptions of more self-serving motives, while lower estimates signify perceptions of more altruistic punishing motives.

### **3.3.3 The Influence of Contextual Factors on Punishers' Formidability Reputations**

In contrast to a punisher's prosocial reputation, motivation attribution and perceived justness of punishment only significantly influenced perceptions of competence (motivation:  $b = -0.14$ , 95% CI [-0.20, -0.09],  $p < .0001$ ; justness:  $b = 0.36$ , 95% CI [0.30, 0.42],  $p < .0001$ ), but not toughness (motivation:  $b = -0.06$ , 95% CI [-0.12, -0.00],  $p = .04$ ; justness:  $b = 0.07$ , 95% CI [0.00, 0.13],  $p = .046$ ) or dominance (motivation:  $b = 0.02$ , 95% CI [-0.03, 0.08],  $p = .43$ ; justness:  $b = -0.08$ , 95% CI [-0.14, -0.02],  $p = .007$ ).

Attributions of toughness ( $b = -0.12$ , 95% CI [-0.18, -0.06],  $p = .0001$ ) and dominance ( $b = -0.19$ , 95% CI [-0.26, -0.13],  $p < .0001$ ) increased significantly when the punisher was more powerful than the violator, and significantly decreased when they punished a family member rather than a stranger (dominance:  $b = -6.69$ , 95% CI [-11.06, -2.32],  $p = .0027$ ; toughness:  $b = -6.30$ , 95% CI [-10.54, -2.07],  $p = .0035$ ). Perceptions of a punishers' formidability also increased when the violation had been severe (toughness:  $b = 0.12$ , 95% CI [0.06, 0.18],  $p < .0001$ ; dominance:  $b = 0.05$ , 95% CI [-0.01, 0.11],  $p = .1$ ; though only significantly for toughness).

### **3.3.4 The Influence of Contextual Factors on Observers' Behavioural Intentions Towards Punishers**

Punishment context factors influenced participants' behavioural intentions towards punishers similarly to how they affected punishers' prosocial reputations. Observers were significantly more willing to help and associate with punishers when the punishment was deemed justified (help:  $b = 0.59$ , 95% CI [0.52, 0.65],  $p < .0001$ ; associate:  $b = 0.42$ , 95% CI [0.36, 0.48],  $p < .0001$ ) and when punishing motives were perceived as more altruistic (help:  $b = -0.16$ , 95% CI [-0.22, -0.10],  $p < .0001$ ; associate:  $b = -0.13$ , 95% CI [-0.18, -0.07],  $p < .0001$ ). Participants

also reported a higher willingness to help punishers when they punished close others (romantic partner:  $b = 13.81$ , 95% CI [5.69, 21.93],  $p = .0009$ ; family:  $b = 10.86$ , 95% CI [6.24, 15.48],  $p < .0001$ ; friend  $b = 10.77$ , 95% CI [4.17, 17.38],  $p = .001$ ) rather than strangers.

### **3.3.5 An Overview of the Contextual Factors Influencing Punishers' Reputations**

This section consolidates the analysis of contextual factors influencing punishers' reputations. While it may reiterate some findings presented earlier, it provides an alternative angle that may enhance our understanding of how these factors shape a punisher's reputation. This section not only examines the factors' significance across various models but also acknowledges their inclusion in the final top models, even when they do not achieve traditional levels of statistical significance. Despite their statistical non-significance, these factors contribute to the overall explanatory power of the models and may therefore influence the complex dynamics observed. The relative importance of each predictor for every model is further detailed by the sum of weights in Table 3.4. The table illustrates the aggregated impact of each of the seven context factors across the relevant top models, for each of the twelve response variables.

Relevant predictors across the twelve reputation models were how justified the punishment was perceived to be and the punishing motive (significant factors in all models except perceived dominance and toughness), as well as the relationship between punisher and violator (a significant factor in nine of twelve models). Here, punishing strangers versus close others had distinct consequences for a punisher's reputation: while punishers gained a formidable reputation when punishing strangers rather than close others, they were viewed as significantly more prosocial when they punished close others as opposed to strangers. Power dynamics was only a significant factor for toughness and dominance, but was present in all top models besides likeability. When punishers had more power than the target being punished, they were perceived as significantly more formidable. Directionally, there was an equal split between whether punishing someone more or less powerful provided more prosocial reputation benefits (Table 3.2). In addition, violation severity only statistically significantly influenced the perceived toughness of punishers (with punishers being regarded as tougher if they punished a severe violation), although this factor was present in ten out of twelve top models. A variable which may have somewhat (albeit non-significantly) influenced a punisher's reputation - as indicated by its presence in nine of the twelve top models as well as the relative importance of

each predictor for every model as summarised by the sum of weights in Table 3.4 - is violation intentionality. Punishers were directionally seen as more prosocial, and observers directionally had more positive behavioural intentions towards punishers, when an intentional rather than accidental violation was punished. Intentionality was absent from the top models for dominance, toughness, and trustworthiness. Punishing type did not significantly influence perceptions of punishers in any model, but was present in five top models. Directionally, engaging in SPP or a mixed form of punishment bestowed more reputation benefits than engaging in TPP (Table 3.2).

**Table 3.4**

*Importance of punishing context factors across models*

|                    | Justified<br>Punishment | Punishing<br>Motive | Violator<br>Relationship | Violation<br>Intentionality | Power<br>Dynamics | Punishing<br>Type | Violation<br>Severity |
|--------------------|-------------------------|---------------------|--------------------------|-----------------------------|-------------------|-------------------|-----------------------|
| Overall reflection | 1                       | 1                   | -                        | 0.81                        | 0.18              | -                 | 0.22                  |
| Trustworthiness    | 1                       | 1                   | 1                        | -                           | 0.4               | 1                 | 0.17                  |
| Generosity         | 1                       | 1                   | 1                        | 1                           | 0.28              | 0.47              | 0.24                  |
| Reliability        | 1                       | 1                   | 0.87                     | 0.09                        | 0.64              | 0.44              | -                     |
| Cooperativeness    | 1                       | 1                   | 1                        | 1                           | 0.32              | -                 | 0.2                   |
| Morality           | 1                       | 1                   | 1                        | 0.3                         | 0.15              | -                 | 0.17                  |
| Likeability        | 1                       | 1                   | 1                        | 1                           | -                 | -                 | 0.3                   |
| Competence         | 1                       | 1                   | -                        | 0.34                        | 0.55              | 1                 | -                     |
| Toughness          | 0.72                    | 0.75                | 1                        | -                           | 1                 | -                 | 1                     |
| Dominance          | 1                       | 0.22                | 1                        | -                           | 1                 | -                 | 0.66                  |
| Help willingness   | 1                       | 1                   | 1                        | 1                           | 0.77              | -                 | 0.21                  |
| Association        | 1                       | 1                   | -                        | 1                           | 1                 | 0.66              | 0.21                  |

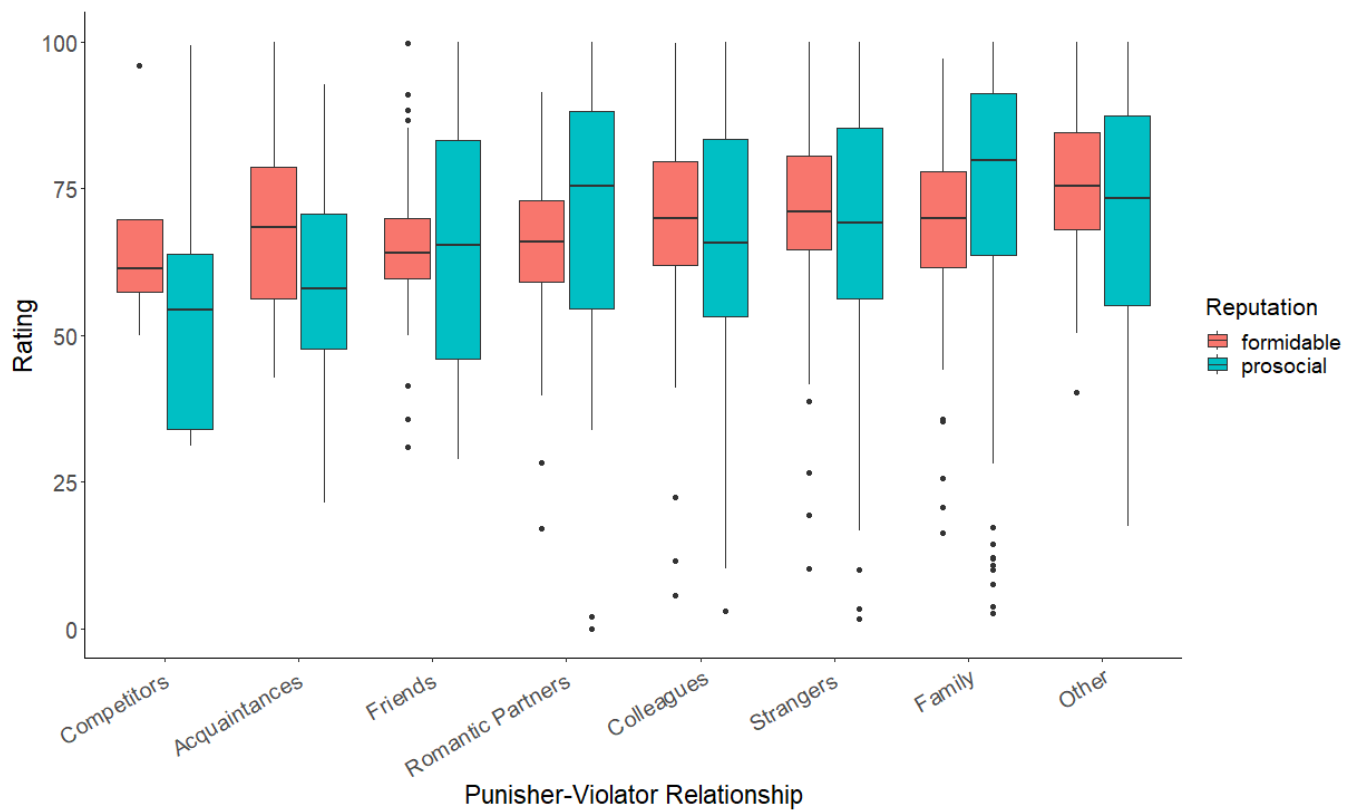
**Note.** This table shows the importance of each of the seven context variables across different models, reflecting their influence on the twelve response variables. The values indicate the aggregated impact of each variable derived from the `sw()` function in the MuMIn package, which calculates the per-variable sum of model weights. Each value represents how strongly the variable contributes to the models' predictions regarding the response variables listed on the left. A value of 1 means the variable was present in all top models and had a strong influence. Values less than 1 indicate varying degrees of importance, with higher values representing greater influence. A dash (-) means the variable was not included in the top models for that response variable.

### 3.3.6 Relationship Effects on Punishers' Formidable versus Prosocial Reputations

To further examine the effects of the relationship between punishers and violators, reputation domain, as well as their interaction, on observers' perceptions of punishers, I conducted a mixed-effects ANOVA. The main effects of violator relationship ( $F(7, 1250.5) = 3.00, p = .004$ ) and reputation domain ( $F(1, 1067.8) = 3.89, p = .049$ ) were statistically significant. Moreover, there was a significant interaction between violator relationship and reputation type,  $F(7, 1067.8) = 4.35, p < .0001$  (Figure 3.2). Comparatively, punishing contributed more to an individual's formidability reputation than to their prosocial reputation when enacted against acquaintances ( $b = 8.79, SE = 3.28, t(1078) = 2.68, p = .008$ ), colleagues ( $b = 4.88, SE = 2.19, t(1078) = 2.23, p = .03$ ), or strangers ( $b = 4.08, SE = 1.88, t(1078) = 2.18, p = .03$ ). Although the punishment of competitors yielded the highest coefficient, the difference was not statistically significant in influencing a punisher's formidable versus prosocial reputation ( $b = 10.33, SE = 10.17, t(1078) = 1.02, p = .31$ ). Punishers gained significantly more prosocial reputation benefits when punishing family members ( $b = -5.31, SE = 1.71, t(1078) = -3.10, p = .002$ ), with the same directional albeit non-significant effect when punishing romantic partners ( $b = -4.34, SE = 4.30, t(1078) = -1.01, p = .313$ ). Overall, the figure shows that formidability reputation benefits were more consistent. On average, punishers received the highest prosocial reputation benefits when punishing family ( $EMM = 73.8, SE = 1.50$ ) or romantic partners ( $EMM = 69.3, SE = 3.39$ ), and the least for punishing acquaintances ( $EMM = 58.60, SE = 2.58$ ) or competitors ( $EMM = 62.5, SE = 7.64$ ). Conversely, formidability ratings were highest on average when punishment was enacted against competitors ( $EMM = 72.9, SE = 7.64$ ), colleagues ( $EMM = 70.60, SE = 1.82$ ) or strangers ( $EMM = 69.8, SE = 1.63$ ), and the lowest for punishing romantic partners ( $EMM = 64.90, SE = 3.39$ ).

#### Figure 3.2

*Comparison of Observer Ratings based on Punisher-Violator Relationships and Reputation Domain*



*Note.* Box plot comparing the distribution of observer ratings across different levels of punisher-violator relationships and reputation domains. Each box represents the IQR of ratings within each category, with the median indicated by the horizontal line inside the box. Dots represent outliers beyond 1.5 times the IQR.

### 3.4 Discussion

The present study used a weekly survey study design in which participants were sampled across a six-week period, to investigate the reputational consequences of punishing. In contrast to the short economic games and specific scenarios frequently used in previous studies (e.g., Dhaliwal et al., 2021; Pedersen et al., 2020; Salcedo García, 2020; Wu et al., 2022; but see e.g., Hofmann et al., 2018; Molho et al., 2020), this study took a broader, more ecologically valid approach (Molho et al., 2020; Uy et al., 2010). This allowed me to test how the reputation of punishers in everyday life varies based on prior predictions and to identify the situational factors that exert the greatest impact on a punisher's reputation. In addition to statistically significant predictors, the discussion also considers factors present in the top models, as these may nevertheless be relevant in somewhat influencing reputational consequences of punishment.

Forming one of my main predictions, Raihani and Bshary (2015b) argued that observers should be sensitive to the apparent motives for punishing, with only punishers who appear motivated by concerns for another's wellbeing rather than by selfish thoughts, receiving reputational benefits. It has frequently been suggested that third-party punishers are likely to receive more reputation benefits than second-party punishers, as the former generally do not stand to directly benefit from punishing and are thereby viewed as other-regarding whilst second-party punishers are often thought to be acting out of self-interest (e.g., Fehr & Gächter 2002; Martin et al., 2019; Raihani & McAuliffe 2012). I additionally specified that any reputation benefits of TPP that SPP lacks, may be limited to cooperative reputations (Kurzban et al., 2007). However, against predictions and prior findings (e.g., Barclay, 2006; Jordan et al., 2016; Jordan & Rand, 2020; Kurzban et al., 2007; Nelissen, 2008; Raihani & Bshary, 2015b, 2015c), I found no evidence to suggest that third-party punishers are evaluated more positively than second-party punishers. Generally, punishers were even evaluated more positively when they engaged in SPP or a mixed form of punishment as opposed to TPP (although this tendency was not statistically significant for any model). Overall, the type of punishment engaged in did not have a strong effect on the reputation of punishers. TPP in itself did therefore not seem to bestow reputational benefits, at least when other factors were considered simultaneously. However, the underlying foundation of prior predictions appeared to be supported: whether a punisher is attributed more selfish or altruistic motives is key to (prosocial) reputation consequences. Motive attribution was a significant predictor in all models besides those predicting perceived dominance and toughness, with punishers receiving prosocial reputation benefits when they were thought to be acting altruistically. Although third-party punishers were viewed as acting significantly more

altruistically than second-party punishers, when punishing type was considered alongside all other factors it did not itself seem to be a strong predictor of how a punisher would be evaluated. In line with this, it has for instance been established that even third-party punishers can be appraised as less interested in fairness restoration and instead desirous of harming a violator when other surrounding factors such as the proportionality of the response are taken into account (Dhaliwal et al., 2021).

Incidentally, two strong predictors of a punisher's reputation were how justified a punishment was perceived to be and the attributed punishing motive, as well as the relationship between punisher and violator. Consistently, punishers received significantly better evaluations when the punishment was considered to be more justified. This finding is in line with prior predictions and theoretical models, which argue that observers should view only justified punishers positively (Barclay, 2006; Panchanathan & Boyd, 2003). Only toughness and dominance were not significantly influenced by justness or motive attributions. Directionally, punishers were even perceived as more dominant when their punishing action was perceived to be less justified and stemming from more selfish motives. It is possible that this (non-significant) effect merely reflects that dominant individuals could be more likely than submissive individuals to punish even when doing so is not justified. Additionally, I found that justified punishers were perceived to be acting significantly more altruistically than unjustified punishers, which is plausible as it would be difficult to view unjustified punishing actions as altruistic.

My findings surrounding the relationship between punisher and violator align with and extend prior studies. These studies have shown that the appropriateness of a punisher's actions is often judged more positively when they are socially connected to the violator, whereas those who punish strangers tend to receive more negative evaluations (Eriksson et al., 2017; Strimling & Eriksson, 2014). In most models, I found that punishing family members, friends, or romantic partners benefitted a punishers reputation more than punishing a stranger. However, the punishers relationship to the violator appeared to yield distinct reputation benefit. Punishers only gained *prosocial* reputation benefits for engaging in punitive actions against close others as opposed to strangers. In contrast, punishers were evaluated as especially strong and dominant when punishing strangers rather than family, thereby gaining reputation benefits in the formidability domain. Further, punishing acquaintances, colleagues, and strangers had a greater impact on the perceived formidability rather than the perceived prosociality of punishers. Only when punishing family members did punishers receive more prosocial than formidable reputation benefits.



Furthermore, punishers were always regarded more positively when the violation had been intentional rather than accidental. However, after Bonferroni corrections this factor was never statistically significant. It is possible, that violation intentionality matters more for prosocial than formidability aspects of a punisher's reputation, as the factor was only absent from top models for toughness and dominance, although any effects are likely weak.

Power dynamics significantly predicted the evaluation of punishers in only two models: punishers were perceived as significantly more dominant and tough when the violator had less power than themselves. Directionally, observers perceived those who punished someone who held more power than themselves as more generous and moral, and indicated both that this reflected more positively on a punisher overall and that they had a greater willingness to help and associate with those punishers in future. On the other hand, when punishers held more power than the violator, they were rated as directionally more trustworthy, reliable, cooperative, and competent. Considering this, it appears that punishers with more power benefit from a formidable reputation, while influences on a punishers prosocial reputation are mixed and weak. Again, these results align with predictions and prior findings. Due to reduced costs of retaliation, punishment is thought to be a more profitable tool for individuals with more power (Gordon & Lea, 2016; Singh & Boomsma, 2015). Further, studies using vignettes have found that powerful punishers are viewed as more formidable (Gordon & Lea, 2016). Although it is plausible that punishing someone with less power makes the punisher seem more dominant and tough, it may also reflect that individuals who hold more power are perceived as tougher and more dominant regardless of punishing behaviours. Furthermore, it is worth noting that previous studies have predominantly focused on formidability reputation aspects of a punisher's reputation when investigating power dynamics. As our findings demonstrate at most directionally weak and mixed influences of power dynamics on a punisher's prosocial reputation, it suggests the possibility of distinct strands of reputational benefits. Power dynamics appear to primarily influence perception of formidability, while the relevance of prosocial reputation aspects may be minimal in this particular context.

The overall weakest predictor of the reputational consequences of punishment was the severity of the violation being punished, which only significantly predicted how tough a punisher would be perceived to be. Generally, punishers received more favourable evaluations (for toughness, trustworthiness, generosity, morality, dominance, the overall reflection on the punisher, as well as willingness to help and associate) for punishing severe violations, but were perceived as directionally more cooperative and likeable when less severe violations were punished. Overall,

this is in line with prior research which found that punishing behaviours are evaluated increasingly positively the more inappropriate a norm violation was (Eriksson et al., 2021), as people believe punishing actions should be proportional to the severity of a violation (Carlsmith et al., 2002; Fehr & Gächter, 2000; Goodwin & Benforado, 2015; Hofmann et al., 2018). Punishment of both more severe and intentional violations may be viewed favourably because they have been found to trigger higher moral outrage (Darley & Pittman, 2003; Reb et al., 2006). It is possible that everyday punishing instances were not severe enough to elicit moral outrage and thereby a stronger effect.

The act of punishing wrongdoings in daily life generally elicited more positive evaluations than negative ones from observers. However, to ascertain whether punishment truly enhances reputation, a comparison to a baseline prior to the actors' punitive behaviours or, alternatively, additional evaluations of non-punishers is necessary. Accordingly, the willingness of observers to associate with punishers did not show a significant increase. This lack of change may also be indicative of the social context. In most reported cases, participants already knew and had formed complex relationships with the punisher and other involved parties. Therefore, everyday instances of punishment, especially small-scale instances with little impact, may not greatly affect an observer's willingness to associate with a punisher they have an established relationship with, or for whom knowledge of past behaviours already exists. In line with this, Macfarlan et al. (2013) for instance demonstrated that while prosocial reputations are associated with cooperative behaviours, they remain relatively consistent over time.

Moreover, punishment in everyday life appears to provide distinct benefits to the reputation domains of prosociality and formidability, particular when considering the relationship between punishers and violators. Future studies could include more distinct facets of an actor's formidability to better investigate the comparative impact of punishment on prosocial and formidability reputations. Similarly, my two measures of behavioural intentions (willingness to help and associate with the punisher) are more in line with a punisher's prosocial reputation. It may therefore be interesting to include other behaviours which are more closely related to one's formidability, such as the likelihood to defect against the punisher or elect them as a leader.

Additionally, it is important to note that several potential biases and confounds may have been introduced to this study, including biases related to participants' reporting tendencies and possible differences in the types of punishment incidents they reported based on their level of involvement—whether they directly experienced the punishment (as a victim or violator), observed it, or learned about it (e.g. via gossip). For instance, participants may have selectively

reported punishment incidents that were memorable, unusual, or otherwise notable. Such incidents may have been easier to recall or deemed more “worthy” of mention, potentially leading to a selective reporting bias both for the participants, and for the individuals who informed participants about recent punishing instances. It is therefore possible that incidents learned about could be more significant or severe than those that were directly observed or experienced, as more notable cases might be more likely to be discussed. In contrast, routine or minor punishments could be underrepresented because they might not be perceived as interesting enough to share. This pattern could skew findings toward more prominent or widely discussed punishment contexts, potentially obscuring insights into more routine, everyday punishments. Although many mundane punishing events were reported, their frequencies were not explicitly examined, and this study did not specifically record whether each punishment incident was directly observed, experienced, or learned about, which future research could address. See Appendix B5 for a sample of participant responses, providing insight into the range of observed, witnessed, or learned-about punishing actions in everyday life.

Additionally, when participants reported on punishment incidents they learned about from others, the credibility of and relationship with the source may have influenced how the incident was perceived and consequently reported. For example, a participant's view of a punishment's justness or a violation's severity may depend on their trust in the informant or any perceived motives behind the information shared. These factors could bias perceptions, as participants might assign different levels of weight or importance based on how they viewed the informant's reliability or intentions. However, this study did not capture details about the sources of gossip.

Furthermore, self-serving biases or concerns about reputation may have decreased participants' likelihood of reporting incidents in which they themselves were the violator. Additionally, when participants were the victim (or were emotionally invested in or interconnected with the victim), they may have been more likely to approve of the punishment or assign higher severity ratings to the incident. Therefore, participant roles and the relationships between individuals should be considered when interpreting the findings. Since participants were uninvolved in the reported punishing instances in most cases (i.e., they were neither victim nor violator), caution should be exercised when drawing inferences about how victims or violators, as opposed to uninvolved individuals, may perceive punishers in everyday life.

Importantly, even with a wide range of known and unknown interaction partners from a multitude of real-life situations in which individuals frequently already have direct information about a punisher's prior behaviour, and without the pre-determined options of the laboratory,

these findings mostly support previous research. With the exception of the type of punishment (TPP, SPP or a combination thereof), every context factor influenced the reputation of punishers (at least directionally) as predicted. Even for the punishing type factor I argued that the underlying premise for why third-party punishers are expected to receive more favourable evaluations than second-party punishers holds true: punishers who appear to be acting out of concern for others are viewed more positively than punishers with seemingly self-serving motives. Altogether, these findings highlight the importance of considering situational factors to gain a comprehensive understanding of the reputational consequences of punishment.

# **Chapter 4**

## **Meta-Cognitive Accuracy in Assessing the Influence of Cooperative and Uncooperative Actions on Reputation**

## 4.1 Introduction

In the complexity of human social interactions, the interplay between self-perception and the perceptions of others holds a pivotal role. Central to this dynamic are meta-perceptions—our beliefs about how others perceive us—and the degree to which these meta-perceptions align with reality, known as meta-accuracy (Kenny & DePaulo, 1993). Meta-perceptions shape our attitudes, guide our behaviours (Byron & Landis, 2020; Kleinlogel et al., 2020; Moore-Berg et al., 2020; Schlenker & Weigold, 1992; Tissera et al., 2021) and are used to initiate, maintain and monitor our relationships (Carlson, 2016; Leary et al., 1995; Srivastava & Beer, 2005). If, for instance, an individual believes their boss views them as highly competent they may pursue a promotion. If, however, they believe their boss sees them as incompetent, they may focus on proving their value first (Byron & Landis, 2020). Inaccurate meta-perceptions in such scenarios could result in embarrassing rejections or missed opportunities (for adverse social outcomes of inaccurate meta-perceptions see Ames & Wazlawek, 2014; Anderson et al., 2008; Brion & Anderson, 2013). Accurate meta-perceptions are therefore essential for employing appropriate interaction strategies (Brion et al., 2015) and are associated with improved relationship quality (Carlson, 2016). People even favour relationships with individuals who possess accurate meta-perceptions irrespective of whether those perceptions are positive or negative (Carlson, 2016). In other words, positive relationships and effective reputation management rely on the accuracy of an individual's meta-perceptions.

Although individuals sometimes overestimate how well others can detect their internal states (e.g. Cameron & Vorauer, 2008; Kaplan et al., 2009), they are aware that other individuals do not see them as they see themselves (Carlson & Kenny, 2012; Carlson et al., 2011), and are fairly accurate at estimating how others – both strangers and acquaintances - perceive them (Carlson et al., 2010; Carlson et al., 2011; Kenny & DePaulo, 1993; Shechtman & Kenny, 1994). Moreover, individuals can differentiate how they are perceived across social contexts (Carlson & Furr, 2009), and can even predict how others will view them when their attempts to manage their reputation are unsuccessful (Wu & Zheng, 2019). However, most research on perception accuracy concern stable personality traits. There is limited research on whether individuals are aware of the motives others attribute to them and how others perceive them based on specific actions. To further our understanding on reputation and reputation management in everyday life, I conducted an online study to investigate whether we know how our cooperative and uncooperative actions influence motive attributions and our reputation (for pilot studies see Appendix C1).

It is worth considering that people may be better at predicting how they will be evaluated for their cooperative actions than how they will be perceived based on their uncooperative actions. Cooperative actions are generally more aligned with social norms, therefore more practiced, and more conducive to positive reputation management than uncooperative actions.

The feedback we receive from others can influence the accuracy of our assessments, especially if we do not correct for potential biases (Fay et al., 2012). Polite lies, with which we for instance pretend not to notice the mistakes of an interaction partner, are common (Camden et al., 1984; Turner et al., 1975), and contentious thoughts are veiled with politeness (Brown & Levinson, 1987). Consequently, positive feedback is provided more often than negative feedback in everyday life, and individuals believe that their negative feedback is less successful (Gallrein et al., 2019). Moreover, negative feedback is generally only shared when it is not perceived as a threat to the relationship with the recipient (Gallrein et al., 2019). As assessments of others' opinions are led by observable cues (Funder, 2001; Watson et al., 2000), this bias in feedback sharing may lead to a better understanding of how our cooperative versus uncooperative behaviours are perceived: as cooperative behaviours are more aligned with social expectations, people may be more familiar with the positive reactions and evaluations others give for those behaviours, making it easier for individuals to predict how others will evaluate those actions. It is also possible that fear of negative evaluations or social consequences discourages individuals from openly displaying selfish behaviours, further limiting the opportunities for feedback. Self-reflection on uncooperative behaviours may also be more difficult, as people are motivated to perceive themselves as moral (Fiske, 2014; Leary, 2007).

Not all traits are judged with the same level of accuracy (Letzring & Funder, 2018). Meta-accuracy tends to be higher when traits are observable, i.e. by being linked to more external cues or behaviours (e.g. extroversion), as opposed to being more internal or evaluative (e.g. intelligence or impulsiveness; Carlson & Kenny, 2012; Stopfer et al., 2014). Accuracy also increases when cues are relevant to traits, or actors were observed in situations in which cues for a trait were made available to others (Hirschmüller et al., 2015; Letzring & Funder, 2018). Similarly, when observers have little or not particularly relevant information from which to base their judgements of others, they often rely on either stereotypes or on their view of themselves, rooted in assumed similarity (Beer & Watson, 2008; Cronbach, 1955; Letzring & Funder, 2018; Ready et al., 2000; Paunonen & Hong, 2013; Watson et al., 2000).

Even when actors exhibit high meta-perceptive accuracy, they can display meta-bias. Meta-bias is generally operationalised as the mean-level difference between meta-perceptions and actual

perceptions. People can either err by (i) overestimating how favourably others think of them (i.e. they think others see them more favourably than they actually do), or (ii) underestimate how favourably others view them (i.e. they think others see them less favourably than they actually do). As one cannot eliminate both errors at the same time, making fewer errors of one type results in more errors of the other. According to error management theory (Haselton, 2007; Haselton & Buss, 2000, 2003; Haselton & Nettle, 2006; Haselton et al., 2009), whenever two errors have different cost magnitudes, a strategy which reduces the more expensive error type and increases the cheaper type should be favoured (Kiyonari et al., 2000; Yamagishi et al., 2007). In other words, when facing asymmetric error costs, individuals should be more likely to commit cheap errors to reduce the likelihood of engaging in expensive errors (Efferson et al., 2020; McKay & Efferson, 2010). In the context of reputation, erring on the side of caution would likely involve systematically underestimating how positively others perceive us, as this approach would keep individuals driven to continuously monitor and control their conduct and reputation. Not doing so when one ought to can bring serious consequences, such as permanent harm to relationships or social isolation (FeldmanHall & Shenhav, 2019). Indeed, previous research has found that uncertainty about payoff consequences of cooperative behaviours causes people to use social heuristics which lead to an increase in cooperation (van den Berg et al., 2020), and that judges tend to like and respect actors less when these actors have a more favourable view of themselves than the observer does (Barranti et al., 2016). Correspondingly, recent research has demonstrated a negative meta-bias on personality attributes (Elsaadawy & Carlson, 2022). Similarly, even from the age of five, people tend to underestimate how much their conversation partner likes them (Boothby et al., 2018; Mastroianni et al., 2021; Wolf et al., 2021). However, these finding may be specific to conversations (Welker et al., 2023). During conversations, individuals may be over-focusing on small social blunders or cues of rejection (Gilovich et al., 2000). Searching for negative evaluations during these interactions may make negative thoughts more salient, leading to pessimistic meta-perceptions (Boothby et al., 2018). It is therefore less certain how biased meta-perceptions might be in settings that do not involve individuals receiving feedback from the evaluator – especially when both actor and evaluator are anonymous and never interact with each other. Indeed, one online study revealed that immoral actors systematically overestimated the extent of positive motive attributions (Lees et al., 2022), and another showed that individuals believed they appeared warmer in their selfies but less competent in their text than they did in reality (Tsankova & Tair, 2022).



Attributes connected to warmth (Anderson, 1968; Wojciszke et al., 1998), especially when related to moral traits (Brambilla et al., 2011; Goodwin et al., 2014), can leave particularly lasting impressions on others. Consequently, this study placed particular emphasis on moral attributes and motives. People often ascribe moral characteristics to others, even in the early stages of interactions and with limited available information (Goodwin et al., 2014; Todorov et al., 2008). Moral traits are of particular interest because they provide insights into whether someone is likely to be a desirable or undesirable interaction partner (Abele & Wojciszke, 2007; Cottrell et al., 2007; Leach et al., 2007; Strohminger & Nichols, 2014; Wojciszke et al., 1998a, b).

Similarly, people frequently and instinctively infer others' motives (Aarts et al., 2004; Dik & Aarts, 2007; Hassin et al., 2005; Malle & Holbrook, 2012; Moskowitz & Olcaysoy Okten, 2016) from a young age (Baillargeon et al., 2015; Gergely et al., 1995; Hamlin et al., 2007; Liu et al., 2007). Understanding the motivations behind people's actions is important, as these motives expose their character and, consequently, forecast their subsequent behaviours (Davis, 1983; Goodwin, 2015; Buckels et al., 2013). Furthermore, motives are crucial for connecting behaviours to judgements of character, as different motivations can drive the same actions (Reeder, 2009). Helping behaviours can for instance be driven by empathy or notions of fairness (Batson, 2011), the desire to signal virtue (Ariely et al., 2009; Berman et al., 2015; Kraft-Todd et al., 2020), or to avert feelings of guilt (Charness & Dufwenberg, 2006). Individuals also harm, punish or deceive others both for selfish (Crocket et al., 2014; Marshall et al., 2021; Shalvi et al., 2015) and other-oriented reasons (Erat & Gneezy, 2012; Fehr & Gächter, 2002; Levine & Schweitzer, 2015; West et al., 2022). Individuals, however, often assume that others' behaviours are motivated by self-interest (Brick et al., 2021; Epley & Dunning, 2000; Kruger & Gilovich, 1999; Miller, 1999; Miller & Ratner, 1998; Ratner & Miller, 2001).

There is evidence to suggest that observers are sensitive to the potential motives held by actors, and actors themselves possess an awareness of how their actions might mirror their underlying motives. For instance, observers tend to hold a more favourable view of non-helpers compared to helpers driven by selfish motives (Alcala et al., 2022; Berman & Silver, 2022; Carlson & Zaki, 2018; Newman & Cain, 2014; Raihani & Power, 2022). Conversely, actors often strive to signal their altruistic motives by either forgoing rewards for engaging in prosocial actions (Kirgios et al., 2020) or by engaging in more costly prosocial actions (Olivola & Shafir, 2013). Furthermore, uncooperative behaviours are generally perceived as more indicative of moral traits than cooperative behaviours, largely because individuals often harbour hidden motives

for their cooperative actions, such as building a favourable reputation (Reeder & Brewer, 1979). Extreme or uncommon behaviours are also perceived as particularly informative regarding someone's character (Ditto & Jemmott, 1989; Fiske, 1980; Kelley, 1967; McKenzie & Mikkelsen, 2007). However, the extent to which observers accurately infer actors' motives and actors' predictive abilities regarding how observers perceive their motives, grounded in cooperative or uncooperative actions, is relatively unexplored.

To investigate whether we know how our cooperative and uncooperative behaviours influence our reputation, I used an adaptation of the Social Accuracy Model (SAM; Biesanz, 2010; Lees et al., 2022). With this componential approach I was able to investigate whether actors can accurately predict how they are perceived and misperceived. Besides meta-accuracy (whether we know the relative impression we make on others), I also looked for meta-bias (systematic biases in meta-judgement). To do so, I solicited written descriptions of recent cooperative and uncooperative behaviours with accompanying self- and meta-perceptions. The written accounts of recent behaviours were subsequently provided to a separate set of observers. This design enabled me to directly test actors' meta-accuracy, as well as the extent to which observers can predict the self-reported motives and traits of actors in the context of real-world cooperative and uncooperative behaviours.

Based on the previous findings outlined above, I formed several hypotheses. Firstly, I anticipated that participants would provide accurate meta-perceptions and observer judgments for both cooperative and uncooperative behaviours. However, I expected that meta-accuracy and observer judgments would be lower for uncooperative actions when compared to cooperative ones. When actors misestimate, I expected them to predict others would judge them less favourably than they actually do, indicative of a negative meta-bias. Furthermore, I predicted that participants would demonstrate higher meta-accuracy for relevant, observable items (e.g. cooperativeness) compared to internal (e.g. motivation items) or less relevant items (e.g. competence). Additionally, I anticipated that both actors and judges would perceive accounts of uncooperative behaviour as more indicative of actor character than accounts of cooperative behaviour, but that they would expect actor predictions and actual observer judgements to align more for cooperative compared to uncooperative behaviours.

Finally, I expected to reveal several trait moderators of meta-accuracy. Firstly, I expected participants who place themselves higher on the subjective social status ladder to demonstrate lower meta-accuracy than participants ranking lower on the social status ladder. Meta-perceptive accuracy may be especially useful for individuals of lower-status, as it is especially

important for lower-status individuals to anticipate outcomes which higher-status individuals control (Fiske, 1993; Snodgrass et al., 1998). Furthermore, individuals of higher status tend to pay less attention to lower-status individuals (Goodwin et al., 2000), which could increase the likelihood of those individuals missing cues necessary for accurate meta-perception. Indeed, prior research has shown that individuals of lower status tend to have higher meta-accuracy than higher-status individuals (Carlson, 2016). However, another recent study found that people with low social economic status (SES) incorrectly believed that they are viewed more negatively than individuals with high SES (Engstrom et al., 2023).

In addition, I anticipated that concern for reputation may positively moderate meta-accuracy. Individuals who desire to have a good reputation might pay more attention to observers. Such an active awareness and motivation to understand perceivers could result in greater meta-accuracy.

Similarly, I expected those with a higher propensity to engage in social perspective taking to exhibit enhanced meta-perceptive accuracy. Tsankova & Tair (2021) suggest that a positive association between social competencies and meta-accuracy is possible, as there are reports of impaired meta-accuracy in autism (Sasson et al., 2018). Social perspective taking skills may therefore be necessary for individuals to accurately predict how others will view them.

The preregistration for this chapter can be found in Appendix C2.

## 4.2 Methods

### 4.2.1 Sampling Plan

I derived my sample size from Lees et al. (2022), who examined how actors think others perceive their morally questionable behaviour using similar methodologies to the present study. Their fourth study – the study which is most similar to the present study - produced a significant moderator of meta-accuracy (my most complex relevant model). Using the "mixedpower" package in R I simulated power for the significant moderator of the meta-accuracy model. This data-based estimation used the beta coefficients found in the Study 4 data. Actor participant was selected as the random effect to vary in the simulation, with steps between 60-150 as the different sample sizes power should be simulated for. The critical  $t$ -value was set to 2, which reflects an alpha level of 5%. The results suggested that 140 actors would be sufficient to detect a moderating effect of meta-accuracy of similar size (simulated power = 80%, based on 1000 simulations). I therefore aimed to collect 140 usable cooperative behaviours descriptions and 140 usable uncooperative behaviour descriptions. Further, I planned to oversample by 40 participants, anticipating the need for exclusions, resulting in an aim to sample 180 actor participants.

Subsequently, the sample size for observer participants was based on the decision to have at least 10 observers rate each behaviour description, and an additional trade-off between a high number of observations and time (i.e., to reduce attrition and attention loss I aimed to limit the survey time to 15 minutes). Each observer was therefore asked to rate 8 stories, meaning that 350 observers were needed so that each of the anticipated 280 descriptions (140 for each behaviour) would be rated by around 10 observers. In anticipation of exclusions, I aimed to sample 380 observer participants.

### 4.2.2 Participants and Design

609 participants were recruited via Prolific Academic in two waves and were paid at a rate of £9 per hour. Only those who were fluent in English and had an approval score of 100 were invited to take part in these studies. First, 181 actors (note: 180 were preregistered, but an additional participant who had completed the study was in the 'timed out' section on Prolific) were asked to describe a recent cooperative and a recent uncooperative behaviour, before providing self- and meta-perceptions. Responses from 22 actors were excluded from further

study for not describing a cooperative/uncooperative behaviour, not responding in English, or for being unclear (e.g., due to containing too little information, such as in the responses “I did a favor for a friend” and “Behaving cooperatively for me is when i try helping someone with a real need”). This left 159 actor participants (78 male, 81 female, 2 preferred not to say) between the ages of 18-66 years ( $M = 29.18$ ,  $SD = 7.33$ ) from 28 different countries. Most participants from this first stage of data collection were South African (31%), Portuguese (13%), Polish (10%), British (6%), or Italian (6%). Consequently, 318 behaviour descriptions from 159 actor participants were presented to 428 observer participants, who each read and judged a random selection of 8 descriptions. Three observers failed at least one of two attention checks and were therefore removed from all analyses. The remaining observers (231 male, 187 female, 5 ‘other’, 2 preferred not to say) were between 18-66 years old ( $M = 34.81$ ,  $SD = 11.50$ ) and from 42 different nations (e.g., 34% British, 13% US American, 10% South African, and 6% Polish). After exclusions ( $n = 3$ ), eight stories were rated by nine observers, 82 by ten and 228 by eleven observers ( $Md = 11$ ,  $M = 10.69$ ,  $SD = 0.51$ ). Actors were not able to participate again as observers. The median completion time was 20 minutes 44 seconds for actors and 17 minutes 13 seconds for observers. All participants provided informed consent and the research was approved by the UCL Research Ethics Committee (Project ID: *ICN-PWB-NR-23-10-20A*).

### **4.2.3 Procedure**

#### ***Stage 1 of Data Collection - Actor Participants***

After providing general informed consent, participants from the first stage of data collection received further information before being asked for permission to use their descriptions as stimuli in future research. Specifically, they were told that the study aimed to investigate whether people can accurately predict how others will judge them based on their real-life behaviours. Participants were informed that they would take part in the first stage of this study, in which they would be asked to describe a recent cooperative and a recent uncooperative behaviour, and to predict how others would judge them based on that description. Further, it was explained that in the second stage of data collection, another set of participants would read their behaviour descriptions and rate how they view the participant based on the behaviour. Participants who did not want their descriptions to be used as stimuli in future research were asked to return the survey on Prolific. The remaining participants were asked to provide a written description of a recent cooperative or uncooperative behaviour – the order of whether

they were asked to describe a cooperative or uncooperative behaviour first was counterbalanced. The prompt for both behaviours read as follows: *“In our daily lives we sometimes behave cooperatively or uncooperatively, for various reasons. Examples of uncooperative behaviours include selfish actions such as not pulling your weight, littering, or being unhelpful. Examples of cooperative behaviours include prosocial actions such as helping or collaborating well with others, donating blood, and sharing. Please describe in 2-5 sentences, one recent instance in which you behaved uncooperatively [cooperatively]. After providing your description, you will be asked to provide self-perceptions and predictions for how others will evaluate you for this behaviour. Please do not include any identifying information in your writing. However, please be clear and specific enough so that future participants will understand your description. Providing context is encouraged.”* It was also noted that future participants would see the behaviour exactly as they described it, without any changes made. The behaviour descriptions also acted as a comprehension and attention check.

The next page displayed the exact description the participant had given, along with a request to provide self- or meta-perceptions. Whether actors were first asked to provide self- or meta-perceptions was counterbalanced, as was the order of items within these sections. The prompt for the provision of meta-perceptions encouraged participants to imagine that someone else – the average person- would read their described behaviour and decide how they view them based on it. Actors then indicated how they thought others would rate their traits and motives based on their described behaviour. For self-perceptions actors were asked to indicate what motivated their described behaviour and how they would rate their own traits based on it. The same motives and traits were included for both cooperative and uncooperative behaviours. Therefore, positively ( $n = 4$ ), negatively ( $n = 5$ ) and more ambiguously or neutrally ( $n = 6$ ) valenced motive items, which encompassed both self- and other oriented motives, were included. These items included “self-interest” and “selflessness” (Barasch et al., 2014; Brañas-Garza et al., 2014; Crocket et al., 2014; Levine & Schweitzer, 2015), along with “gaining a good / avoiding a bad reputation”. Further items were: “avoiding conflict” and “helping others”, reflecting approach and avoidance motives (Janoff-Bulman et al., 2008); “fairness”, reflecting relationship regulation motives (Batson, 2011; Rai & Fiske, 2011); “empathy”, “pride”, and “averting guilt” reflecting moral emotions (Batson, 2011; Charness & Dufwenberg, 2006; Tangney et al., 2007; Teper et al., 2015); as well as “obligation” (Gerstenberg et al., 2018), “spite”, “jealousy”, “competitiveness”, “revenge” (Bone & Raihani, 2015; Brañas-Garza et al., 2014; Deutchman et al., 2021; Pedersen et al., 2013; Twardawski et al., 2022), and “laziness”. I moreover selected a range of reoccurring and relevant moral character, warmth and competence traits:

“trustworthy”, “sincere”, “honest”, “loyal”, “just”, “reliable”, “empathic”, “likeable”, “competent”, and “intelligent” (Abele et al., 2008; Brambilla et al., 2011; Clausell & Fiske, 2005; Cottrell et al., 2007; Cuddy et al., 2008; Fiske et al., 2002; Goodwin, 2015; Goodwin et al., 2014; Judd et al., 2005; Kervyn et al., 2012; Lapsley & Lasky, 2001; Leach et al., 2007; Lin et al., 2005; Uhlmann et al., 2015; Walker & Hennig, 2004; Wojciszke, 2005). In total, actors rated 25 items (15 motives and 10 traits) for each behaviour in both the self-perception and meta-perception conditions. All items were rated from 1 = “*Not at all*” to 7 = “*Completely*”.

The same procedure was then repeated for the remaining behaviour. After behaviour descriptions and ratings had been provided, participants indicated whether they thought they were more accurate in predicting how their cooperative or their uncooperative actions would be judged (rated from 1 = “*Definitely Uncooperative*” to 7 = “*Definitely Cooperative*”, with 4 marked as “*Equal*”) as a meta-meta-accuracy measure. Similarly, they indicated whether they believed the uncooperative or the cooperative behaviour they described was more reflective of their character (rated from 1 = “*Uncooperative Behaviour*” to 7 = “*Cooperative Behaviour*”, with 4 again marked as “*Equal*”). An attention check using the same scale instructed participants to select “*uncooperative behaviour, the very first button*”, to show that they were paying attention.

In random order, participants from the first wave of data collection furthermore completed the neuroticism scale of the Eysenck Personality Questionnaire-Revised Short Scale (EPQR-S) ( $\alpha = 0.82$ ) (Eysenck et al., 1985), placed themselves on the MacArthur Scale of Subjective Social Status ( $M = 5.88$ ,  $SD = 1.51$ ) (Adler et al., 2000), completed the Single-Item Narcissism Scale ( $M = 2.26$ ,  $SD = 1.41$ ) (Konrath et al., 2014), the Social Perspective Taking Scale ( $\alpha = 0.86$ ) (Gehlbach et al., 2008), and the 7-item Concern for Reputation Scale ( $\alpha = 0.84$ ) (De Cremer & Tyler, 2005). All Cronbach’s alphas ( $\alpha$ ) reported in this chapter indicate the internal consistency of each scale based on the data collected in this study. Lastly, they provided demographic information, which completed the first stage of data collection. As two actor participants reported their status as 6.5 and 7.5 respectively, these two values were rounded up.

### ***Stage 2 of Data Collection – Observer Participants***

The second stage of data collection took the 318 usable behaviour descriptions provided by 159 actor participants in the first stage of data collection, and provided them as stimuli for observers to rate. Each behaviour description was rated by around 11 observers.

Matching the language of the first survey, it was explained that in a first stage of data collection previous participants had been asked to describe a recent cooperative and uncooperative behaviour they had engaged in, and that these participants made predictions about how others would view them based on those descriptions. Observers were told that they were now invited to take part in the second stage of the study, in which they would be reading eight of those descriptions, based on which they should indicate how they perceive the authors. Subsequently, they read the same text about cooperative and uncooperative behaviours in daily life as actor participants did. However, the task prompt read “*Below is a description of a cooperative [uncooperative] behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, indicate how you perceive the author [indicate what you believe the author’s motives were]*”. Observer participants read and rated either four cooperative or four uncooperative behaviour descriptions, before reading four randomly selected descriptions of the other behaviour. All behaviour descriptions were evenly presented. Again, item order was randomised. The behaviour descriptions were not altered in any way, and were presented on the same page as the items to be rated. The trait and motive items were identical to those presented to actor participants in the first stage of data collection. After reading and judging the behaviours, observers were asked whether they thought their judgments and previous participants’ predictions were more aligned for cooperative or uncooperative behaviours (again rated from 1 = “*Definitely Uncooperative*” to 7 = “*Definitely Cooperative*”, with 4 = “*Equal*”), and whether they believe the uncooperative or the cooperative behaviours described were more indicative of people’s characters (rated from 1 = “*Uncooperative Behaviour*” to 7 = “*Cooperative Behaviour*”, with 4 = “*Equal*”). An attention check using the same scale asked observers to select “*uncooperative behaviours, the very first button*”, to show that they were paying attention. Subsequently, observers completed another attention check which asked what topic the descriptions had been on. The correct answer was “*cooperative / uncooperative behaviours*”, whilst the incorrect options were “*holiday destinations / staycations*”, “*dietary preferences / allergies*” and “*consumer behaviour / customer manipulation*”. Finally, they provided demographic information.

The exact survey content for actors and observers can be found in Appendix C3.

#### **4.2.4 Analysis Plan**



All analyses were conducted in R (R Core Team, 2023), and most analyses used an adaptation of the social accuracy model framework (Biesanz, 2010) as presented in Lees et al. (2022). With these models, I could directly test for meta-accuracy and observer accuracy by comparing actor participant ratings to observer participant ratings (models detailed in Table 4.1).

The Social Accuracy Model is a componential model of interpersonal perception designed to estimate perceiver, target, dyadic, and other effects for different elements of accuracy (Biesanz, 2010, 2021). It allows researchers to decompose accuracy into its constituent component (most notably normative and distinctive accuracy) and estimate individual differences in these components, as well as biases and moderators (Biesanz, 2010; 2021). Normative accuracy assesses how accurately perceivers understand the average target's personality, reflecting stereotype-based knowledge, while distinctive accuracy measures the ability to recognise the *unique* traits of specific targets (Cronbach, 1955; Furr, 2008; Biesanz, 2010, 2021; Biesanz et al., 2007).

Conceptually, the SAM thus represents an integration of Cronbach's (1955) componential approach with Kenny's (1994; Kenny & La Voie, 1984) Social Relations Model. The SAM also builds on the epistemological framework outlined by Funder's Realistic Accuracy Model (1995, 1999), where the assessment of accuracy is equated with determining the validity of a measured construct (Cronbach & Meehl, 1955; Messick, 1989). This involves evaluating the degree to which perceivers' impressions align with the measured validity criteria, such as observer perceptions compared to self-reports. Unlike other models that focus on how one sees others or is seen by others on a specific trait, the SAM examines how accurately one perceives others and is perceived by others *across* traits (Biesanz, 2010).

Lees et al. (2022) adapted the SAM to suit their paradigm, and I will subsequently outline their approach in more detail, as I followed their method. While the SAM has generally been applied to classic half-block (one group rates another group) and round-robin (everyone rates everyone else) designs, Lees et al. (2022) used a design in which observers rated multiple actors, but actors estimated how they would be perceived on average. This necessitated modifications to the modelling of random effects. Additionally, the original SAM emphasises 'perceptive accuracy' (how accurately individuals perceive others) and 'expressive accuracy' (how accurately individuals are perceived by others). While Lees et al. (2022) also focused on observer perceptive accuracy, they extended the SAM to examine meta-perceptive accuracy - a key measure of interest in the present study. In applying the adapted SAM by Lees et al. (2022), I followed their detailed model specifications as outlined in their supplementary

materials. These specifications (including baseline, distinctive, insight, moderator, and bias models), closely align with Biesanz's (2010) foundational framework. For detailed equations and centring procedures, see the supplementary materials provided by Lees et al. (2022). For the original formulations and theoretical basis of the SAM, refer to the main text in Biesanz (2010).

Using mixed-effects modelling, I fit three main models when investigating actor meta-accuracy and observer judgment accuracy. As distinct behaviours, these models were initially fit separately for both cooperative and uncooperative behaviours (for an exception see the 'Moderators of Accuracy' section). In fitting the linear mixed-effects models I adopted a maximal random structures approach (Barr et al., 2013), with random slopes and intercepts for all linear predictors across all random effects. Whenever a singular fit was obtained, I removed the random slope with the smallest standard deviation. Drawing on the SAM, estimates provided by participants (i.e. the meta-perceptions and observer ratings) were modelled as the response variables, and the true values (i.e. the observer ratings and self-perceptions) were modelled as predictors. When predicting actor meta-accuracy a single random effect (actor) was specified, and when predicting observer accuracy two crossed random effects were included (actor and observer). Using the 'lmerTest' package (Kuznetsova et al., 2017) all models were estimated using restricted maximum likelihood estimation and Welch–Satterthwaite degrees of freedom approximation, unless specified otherwise.

Furthermore, I used cluster centring within actor (Enders & Tofighi, 2007), also called person centring (Furr & Funder, 2004), to ensure that meta- and observer-perceptions were anchored to the mean validation criterion within each actor. With this approach meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor. By doing so, a value of zero represents the average point estimate accuracy within actor, making predictor variables uncorrelated with each other and aiding the interpretation of intercept estimates. Consequently, deviations from zero provide insights into average point-estimate inaccuracies.

### ***Accuracy Models***

Refer to Table 4.1 for an overview of key models. First, I fit the *baseline accuracy* models, which tested for a linear relationship between the estimated and true values in actor meta-perceptions (meta-perception ~ observer judgment) and observer attributions (observer judgment ~ self-perception). However, these baseline accuracy models do not control for

potential confounds of judgment accuracy which may inflate accuracy estimates: meta-accuracy depends both on an understanding of how one is uniquely perceived (distinctive meta-accuracy) and an understanding of how the average actor is perceived (*normative accuracy*; Wood & Furr, 2016). The second models therefore tested for *distinctive accuracy* (Biesanz, 2010; Furr, 2008), by modelling the linear relationship between the estimated value and the normative profile on top of the baseline models' linear relationship between the estimated and true values (e.g. meta-perception  $\sim$  observer judgment + normative profile). The normative profile is the (grand mean centred) distribution of the self-perception means across actors, with random slopes modelled within actor. The third models tested for actor and observer *insight* (Carlson, 2016). Actor insight was operationalized as the linear relationship between meta-perceptions and observers' attributions while controlling for actors' self-perceptions (specifically the self-perception of item for actor, mean-centred within actor). As the model controls for self-perception, the relationship between meta-perceptions and actual judgements can be interpreted as actors' understanding of how they will be misperceived (Lees et al., 2022). Conversely, the relationship between meta-perception and self-perception can be interpreted as *transparency bias* – the extent to which actors overestimate how accurately they will be perceived (Filvoich et al., 1998). Again, random slopes for transparency were modelled within actor. Similarly, observer insight was operationalised as the linear relationship between actual judgements and actors' self-perceptions while controlling for actors' meta-perceptions, which, as in Lees et al. (2022), was interpreted as “opaqueness” in actors' motives. Observer insight then reflects the observer's ability to accurately perceive the disparity between an actor's actual motives and traits and the actor's predicted perceptions.

### ***Moderators of Accuracy***

Furthermore, I explored the role of moderators on meta-perceptive accuracy. These moderators were subjective social status, concern for reputation, social perspective taking, item, and condition (i.e., uncooperative versus cooperative behaviours). Condition was also explored as a moderator on observer judgment accuracy. Building on the distinctive accuracy model, I interacted the moderator (i.e., mean-centred status, mean-centred concern for reputation, mean-centred social perspective taking, item, or condition) with distinctive accuracy (and normative accuracy as a control). All predictors were modelled as random slopes within actor (e.g., Meta-Perception  $\sim$  Judgement\*Status + Normative Profile\*Status + (1+Judgement + Normative Profile + Status|Actor)). However, this level of model complexity often led to reductions in random structure complexity. For the models assessing the moderating role of uncooperative

versus cooperative behaviour on meta-perceptive or judgment accuracy the data sets for the two behaviours were combined.

### ***Meta-Bias***

To examine systematic biases in meta-perceptions I variable-centred the data (rather than centring within actor) and interacted item with observer judgment in the baseline meta-accuracy model. Specifically, meta-perceptions were centred on the mean observer judgment of each item, and observers' judgments on the mean self-perception of each item. Values of zero therefore represent average point estimate accuracy by item, meaning that estimates which significantly differ from zero can be interpreted as systematic directional bias.

Although I also preregistered an examination of moderators of meta-bias, I retrospectively determined this was not feasible. Bias scores did not take into account whether items were positively ( $k = 14$ ), negatively ( $k = 5$ ), or more ambiguously ( $k = 6$ ) valenced. Therefore, the preregistered moderator analyses would have only reflected general tendencies to under- or overestimate, rather than capturing valence-specific biases, rendering such an approach unsuitable for the intended purpose.

### ***Secondary Analyses***

Furthermore, to assess whether certain items were seen as distinctly applicable to uncooperative versus cooperative behaviours, I conducted mixed-effects Analyses of Variance (ANOVAs) and performed pairwise comparisons on the estimated marginal means ("emmeans" package; Lenth, 2023) for the interaction between item and condition, while adjusting for multiple comparisons with a Bonferroni correction. Specifically, the models predicted mean-centred self-perceptions, meta-perceptions or observer attributions, with random intercepts for actors (self- and meta-perceptions) or judges (observer attributions) along with random slopes for both item and condition within each level of actor (i.e., the random slopes for both item and condition were allowed to be correlated within each level of actor). However, this level of model complexity again led to reductions in random structure complexity.

Lastly, to examine (i) whether participants anticipated greater alignment between predictions and actual judgements for uncooperative or cooperative behaviours and (ii) whether they considered uncooperative or cooperative behaviours to be more indicative of actor character, ratings for those questions were centred around the scale midpoint of 4. These adjusted ratings were then employed as the response variables in four constant (intercept-only) linear models.

**Table 4.1***Description of main meta-accuracy and actor bias models*

|                           | Question                                                                                                               | Operationalisation                                                                                                                                              | Model                                                                                                               |
|---------------------------|------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Baseline Meta-Accuracy    | Do actors know how they are perceived by others?                                                                       | Linear relationship between meta-perceptions (estimations) and observer-perceptions (true values)                                                               | Meta-Perception ~ Judgment + (1+Judgment Actor)                                                                     |
| Distinctive Meta-Accuracy | Do actors know how they will be uniquely perceived by others, relative to the average actor?                           | Linear relationship between meta- and observer-perceptions, controlling for the true distribution of attributes in the actor population (the normative profile) | Meta-Perception ~ Judgment + Normative Profile + (1+Judgment + Normative Profile  Actor)                            |
| Normative Accuracy        | Do actors know how the average actor is perceived?                                                                     | Linear relationship between meta-perceptions and the normative profile, in the distinctive accuracy model                                                       |                                                                                                                     |
| Meta-Insight              | Do actors know how they will be <i>misperceived</i> ?                                                                  | Linear relationship between meta- and observer-perceptions, controlling for the actor's self-perceptions                                                        | Meta-Perception ~ Judgment + Self-Perception + (1+Judgment + Self-perception  Actor)                                |
| Transparency Bias         | Do actors overestimate how transparent their motives and attributes are to others?                                     | Linear relationship between meta-perceptions and actors' self-perceptions, in the insight model                                                                 |                                                                                                                     |
| Meta Bias                 | Do actors systematically over or underestimate the extent to which they are attributed a specific attribute or motive? | Mean difference between a specific meta-perception and its corresponding observer perception, across all actors                                                 | Meta-Perception ~ Judgment * Item + (1+Judgment Actor)                                                              |
| Meta-Accuracy Moderator   | Is the moderator, e.g. subjective social status, associated with more or less meta-accuracy?                           | Measure which moderates the slope of any accuracy coefficient                                                                                                   | Meta-Perception ~ Judgment * Status + Normative Profile * Status + (1+Judgment + Normative Profile + Status  Actor) |

*Note.* Explanation of each main model used in the analyses, including descriptions of operationalisation and model terms. Contains only actor accuracy and bias for simplicity.

Adapted and extended from a table in Lees et al. (2022).

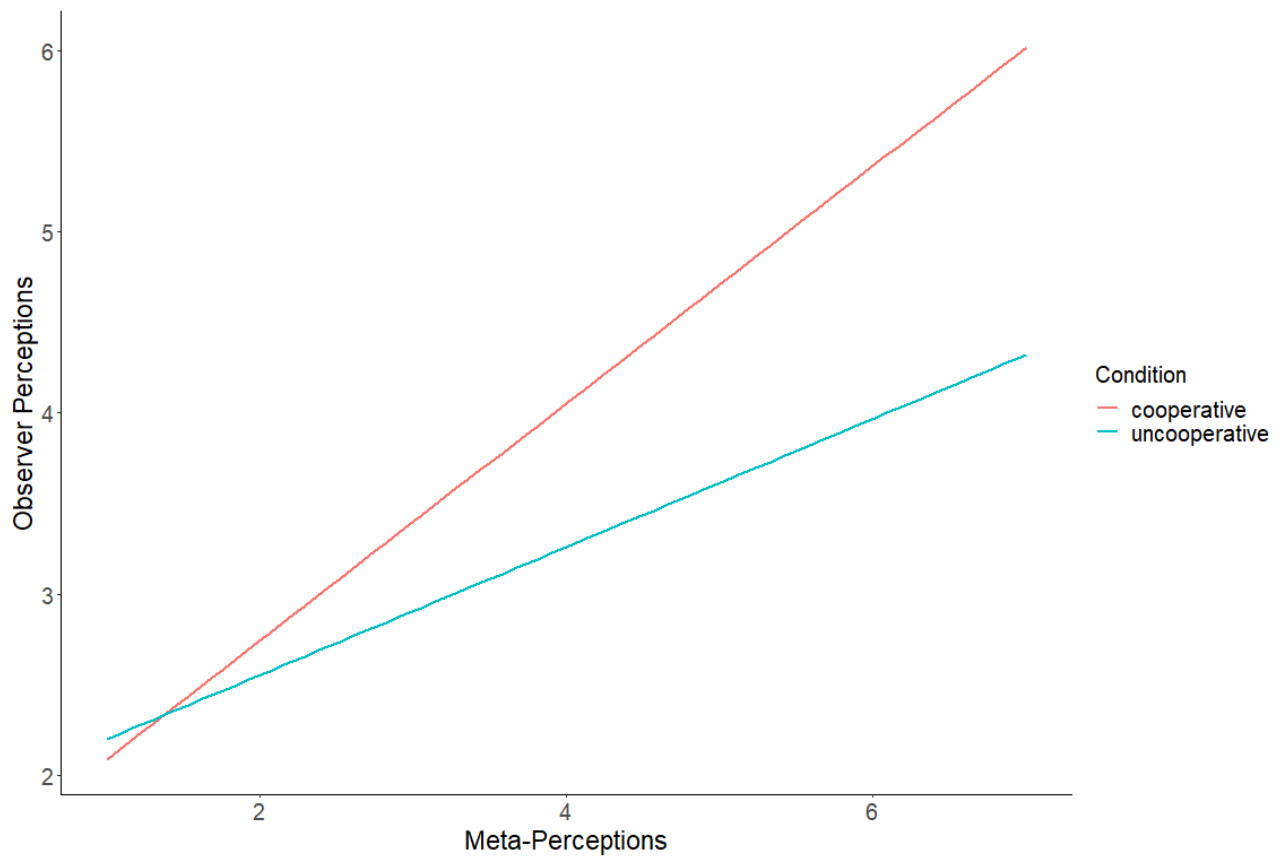
## 4.3 Results

### 4.3.1 Actor Meta-Accuracy

The primary focus of interest was to determine whether actors demonstrate meta-accuracy. In essence, meta-accuracy refers to the congruence between meta-perceptions (how actors expect to be viewed) and observer perceptions (how observers actually view actors). Actors demonstrated baseline meta-accuracy, as evidenced by a significant linear relationship between actor meta-perceptions and observer judgments for both cooperative ( $b = 0.61$ , 95% CI [0.58, 0.64],  $t(157.53) = 43.69$ ,  $p < .0001$ ) and uncooperative behaviours ( $b = 0.32$ , 95% CI [0.29, 0.34],  $t(157.38) = 22.52$ ,  $p < .0001$ ). Figure 4.1 illustrates this alignment, showing that actors' expectations of how they would be perceived closely matched actual observer perceptions, with a particularly strong correspondence for cooperative behaviours. See Table 4.2 (cooperative behaviours) and Table 4.3 (uncooperative behaviours) for details on the three main models fit for each behaviour: baseline accuracy models, distinctive accuracy models, and insight models.

#### Figure 4.1

*Relationship between actor meta-perceptions and actual observer perceptions for cooperative and uncooperative behaviours*



*Note.* The figure presents the raw, uncentered data showing the relationship between actor meta-perceptions and corresponding observer judgments for both cooperative and uncooperative behaviours.

**Table 4.2**

*Actor Meta-Perceptive Accuracy for Cooperative Behaviours Across Models*

| <i>Predictors</i>  | <b>Base Accuracy</b>    |          |          | <b>Distinctive Accuracy</b> |          |          | <b>Insight</b>          |          |          |
|--------------------|-------------------------|----------|----------|-----------------------------|----------|----------|-------------------------|----------|----------|
|                    | <i>b (CI)</i>           | <i>t</i> | <i>p</i> | <i>b (CI)</i>               | <i>t</i> | <i>p</i> | <i>b (CI)</i>           | <i>t</i> | <i>p</i> |
| (Intercept)        | -0.00<br>(-0.14 – 0.13) | -0.03    | .98      | 0.02<br>(-0.08 – 0.12)      | 0.38     | .71      | -0.01<br>(-0.11 – 0.09) | -0.16    | .88      |
| Observer Judgments | 0.61<br>(0.58 – 0.64)   | 43.69    | < .001   | 0.08<br>(0.07 – 0.10)       | 9.33     | < .001   | 0.16<br>(0.14 – 0.18)   | 16.50    | < .001   |
| Normative Profile  |                         |          |          | 0.85<br>(0.80 – 0.90)       | 35.85    | < .001   |                         |          |          |

|                  |                           |           |        |
|------------------|---------------------------|-----------|--------|
| Self-Perceptions | 0.68<br>(0.65 – 0.71<br>) | 42.1<br>8 | < .001 |
|------------------|---------------------------|-----------|--------|

*Note.* This table presents actor meta-perceptive accuracy for cooperative behaviours across the three main model types fit in this chapter. Estimates (*b*), 95% confidence intervals (CIs), *t*-values, and *p*-values are reported. Significant results are bolded. The **baseline accuracy model** tested for a linear relationship between meta-perceptions and observer-perceptions. As meta-accuracy depends both on an understanding of how one is uniquely perceived (*distinctive meta-accuracy*) and an understanding of how the average actor is perceived (*normative accuracy*), the **distinctive accuracy model** tested for a linear relationship between meta-perceptions and the normative profile (the grand mean centred distribution of the self-perception means across actors) on top of the baseline models' linear relationship between meta-perceptions and observer-perceptions. Here, the value for observer judgments reflects distinctive meta-accuracy, while the value for the normative profile reflects normative accuracy. The **insight model** tested for a linear relationship between meta-perceptions and observer-perceptions while controlling for actors' self-perceptions. As this third model controlled for self-perceptions, the relationship between meta-perceptions and observer perceptions can be interpreted as actors' knowledge of how they will be misperceived (*insight*), while the relationship between meta-perception and self-perception can be interpreted as *transparency bias* – the extent to which actors overestimate how accurately they will be perceived. Additionally, as I used person centring (cluster centring within actor) so that meta- and observer-perceptions were anchored to the mean validation criterion within each actor (i.e., meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor), a value of zero represents the average point estimate accuracy within actor, and deviations from zero provide insights into average point-estimate inaccuracies.

**Table 4.3**

*Actor Meta-Perceptive Accuracy for Uncooperative Behaviours Across Models*

| Predictors | Base Accuracy |          |          | Distinctive Accuracy |          |          | Insight       |          |          |
|------------|---------------|----------|----------|----------------------|----------|----------|---------------|----------|----------|
|            | <i>b</i> (CI) | <i>t</i> | <i>p</i> | <i>b</i> (CI)        | <i>t</i> | <i>p</i> | <i>b</i> (CI) | <i>t</i> | <i>p</i> |



|                       |                                |           |                 |                            |           |                 |                            |           |                 |
|-----------------------|--------------------------------|-----------|-----------------|----------------------------|-----------|-----------------|----------------------------|-----------|-----------------|
| (Intercept)           | 0.09<br>(-<br>0.04 – 0.2<br>2) | 1.34      | .18             | 0.07<br>(-<br>0.06 – 0.19) | 1.09      | .28             | 0.07<br>(-<br>0.04 – 0.19) | 1.25      | 0.21            |
| Observer<br>Judgments | 0.32<br>(0.29 – 0.<br>34)      | 22.5<br>2 | < . <b>.001</b> | 0.19<br>(0.17 – 0.22<br>)  | 13.6<br>5 | < . <b>.001</b> | 0.13<br>(0.11 – 0.15<br>)  | 13.1<br>3 | < . <b>.001</b> |
| Normative<br>Profile  |                                |           |                 | 0.67<br>(0.58 – 0.77<br>)  | 14.5<br>6 | < . <b>.001</b> |                            |           |                 |
| Self-<br>Perceptions  |                                |           |                 |                            |           |                 | 0.56<br>(0.52 – 0.60<br>)  | 25.1<br>4 | < . <b>.001</b> |

*Note.* This table presents actor meta-perceptive accuracy for uncooperative behaviours across the three main model types fit in this chapter. Estimates (*b*), 95% confidence intervals (CIs), *t*-values, and *p*-values are reported. Significant results are bolded. The **baseline accuracy model** tested for a linear relationship between meta-perceptions and observer-perceptions. As meta-accuracy depends both on an understanding of how one is uniquely perceived (*distinctive meta-accuracy*) and an understanding of how the average actor is perceived (*normative accuracy*), the **distinctive accuracy model** tested for a linear relationship between meta-perceptions and the normative profile (the grand mean centred distribution of the self-perception means across actors) on top of the baseline models' linear relationship between meta-perceptions and observer-perceptions. Here, the value for observer judgments reflects distinctive meta-accuracy, while the value for the normative profile reflects normative accuracy. The **insight model** tested for a linear relationship between meta-perceptions and observer-perceptions while controlling for actors' self-perceptions. As this third model controlled for self-perceptions, the relationship between meta-perceptions and observer perceptions can be interpreted as actors' knowledge of how they will be misperceived (*insight*), while the relationship between meta-perception and self-perception can be interpreted as *transparency bias* – the extent to which actors overestimate how accurately they will be perceived. Additionally, as I used person centring (cluster centring within actor) so that meta- and observer-perceptions were anchored to the mean validation criterion within each actor (i.e., meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor), a value of zero represents the average point estimate accuracy within actor, and deviations from zero provide insights into average point-estimate inaccuracies.

However, this univariate approach conflates two distinct sources of meta-knowledge: distinctive accuracy (knowledge of how one is uniquely perceived) and normative accuracy (knowledge of how the average target is perceived). The Social Accuracy Model (Biesanz, 2010) and its adaptation by Lees et al. (2022) disaggregate these two components of accuracy.

Thus, distinctive meta-accuracy models then controlled for the average distribution of motives and trait self-perceptions (the normative profile). For both behaviours, I found evidence of distinctive accuracy (cooperative:  $b = 0.08$ , 95% CI [0.07, 0.10],  $t(163.53) = 9.33$ ,  $p < .0001$ ; uncooperative:  $b = 0.19$ , 95% CI [0.17, 0.22],  $t(159.27) = 13.65$ ,  $p < .0001$ ), meaning that actors exhibited meta-accuracy in judging how they are uniquely perceived relative to the average actor describing a cooperative or uncooperative behaviour. Additionally, the normative profile was associated with meta-perceptions for both cooperative ( $b = 0.85$ , 95% CI [0.80, 0.90],  $t(158.63) = 35.85$ ,  $p < .0001$ ) and uncooperative behaviours ( $b = 0.67$ , 95% CI [0.58, 0.77],  $t(158.10) = 14.56$ ,  $p < .0001$ ).

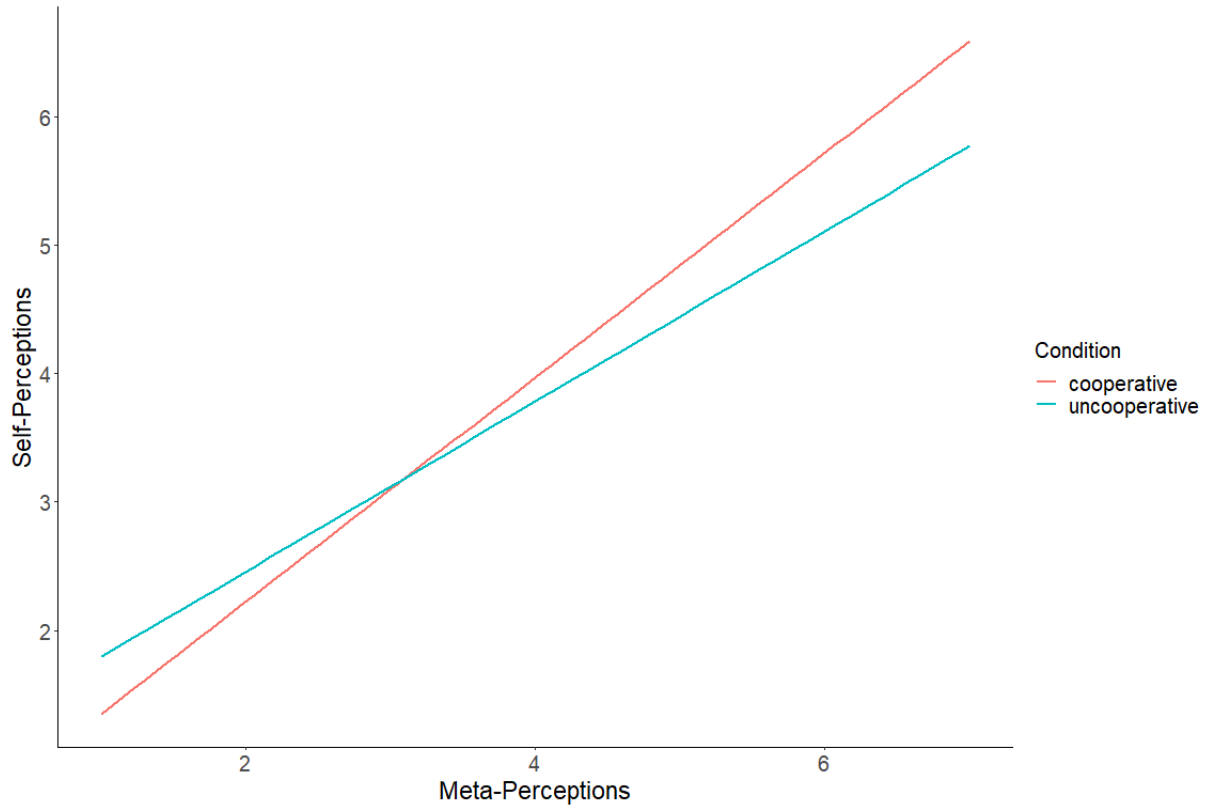
Two additional components of meta-accuracy are insight (actors' awareness of how they are misperceived) and transparency bias (the extent to which actors overestimate how accurately they will be perceived).

Therefore, additional models examined the linear relationship between actor meta-perceptions and observer evaluations (insight) while controlling for actors' actual motives and trait self-perceptions (transparency bias). I found that actors could accurately predict how observers' perceptions of their motives and traits would differ from their actual motives and self-perceived traits, meaning that actors displayed knowledge of how they would be misperceived by observers for their cooperative ( $b = 0.16$ , 95% CI [0.14, 0.18],  $t(158.22) = 16.50$ ,  $p < .0001$ ) and uncooperative behaviours ( $b = 0.13$ , 95% CI [0.11, 0.15],  $t(158.95) = 13.13$ ,  $p < .0001$ ). Simultaneously, actor participants displayed a transparency bias for both cooperative ( $b = 0.68$ , 95% CI [0.65, 0.71],  $t(155.84) = 42.18$ ,  $p < .0001$ ) and uncooperative behaviours ( $b = 0.56$ , 95% CI [0.52, 0.60],  $t(147.78) = 25.14$ ,  $p < .0001$ ), meaning that actors believed their motives would be more apparent to observers than they actually were.

This finding is mirrored in the high correlation between actors' meta- and self-perceptions for both cooperative ( $r(42498) = .84$ ,  $p < .0001$ ) and uncooperative behaviours ( $r(42498) = .64$ ,  $p < .0001$ ), as illustrated in Figure 4.2.

**Figure 4.2**

*Relationship between actors' meta-perceptions and actors' self-perceptions for cooperative and uncooperative behaviours*



*Note.* The figure presents the raw, uncentered data showing the relationship between actor meta-perceptions (how actors thought they would be perceived by others) and actor self-perceptions (how actors viewed themselves) for both cooperative and uncooperative behaviours.

### 4.3.2 Moderators of Meta-Accuracy

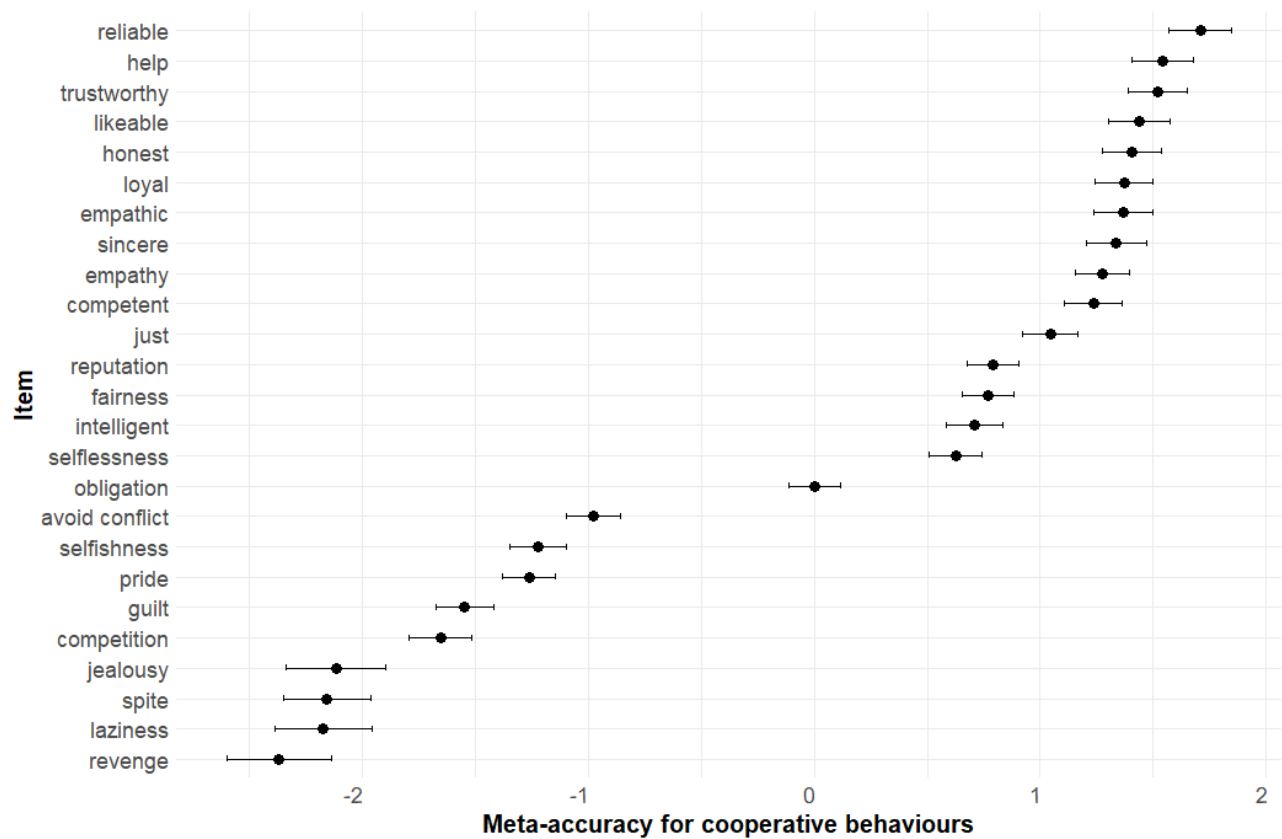
To test for moderators of meta-accuracy, each moderator was separately entered into the distinctive accuracy model as an interaction with observers' evaluations (and as an interaction with the normative profile as a control). In line with the coefficients reported above, behaviour type significantly influenced accuracy: actors exhibited higher distinctive meta-accuracy for uncooperative compared to cooperative behaviours ( $b = 0.12$ , 95% CI [0.10, 0.13],  $t(84767.43) = 18.92$ ,  $p < .0001$ ) and higher normative accuracy for cooperative compared to uncooperative behaviours ( $b = -0.19$ , 95% CI [-0.21, -0.17],  $t(84630.83) = -16.85$ ,  $p < .0001$ ). Social subjective status did not moderate accuracy for either cooperative ( $b = 0.002$ , 95% CI [-0.009, 0.01],  $t(164.41) = 0.39$ ,  $p = .70$ ) or uncooperative behaviours ( $b = 0.02$ , 95% CI [-0.001, 0.04],

$t(157.92) = 1.84, p = .07$ ). Accuracy was also not moderated by reputation concern for cooperative ( $b = 0.002, 95\% \text{ CI } [-0.0007, 0.005], t(163.06) = 1.50, p = .14$ ) or uncooperative behaviours ( $b = 0.0005, 95\% \text{ CI } [-0.004, 0.005], t(158.84) = 0.21, p = .84$ ). The same is true for social perspective taking (cooperative:  $b = 0.003, 95\% \text{ CI } [-0.0008, 0.006], t(159.34) = 1.50, p = .14$ ; uncooperative:  $b = 0.004, 95\% \text{ CI } [-0.001, 0.01], t(125.94) = 1.56, p = .12$ ).

Due to model constraints, item had to be interacted with observer evaluations in the baseline accuracy models. Item was a significant moderator of meta-accuracy for both cooperative ( $F(24, 42169) = 15.00, p < .0001$ ) and uncooperative behaviours ( $F(24, 42198) = 67.37, p < .0001$ ). Estimated marginal means based on the interaction between item and observer judgement, visualised in Figure 4.3, indicated that for cooperative behaviours, actors were most accurate in predicting how they would be perceived on the traits of reliability, trustworthiness, likeability, and honesty, as well as the motive of helping others. Conversely, they were least accurate in predicting the motive attributions of revenge, laziness, spite, jealousy, and competition (Figure 4.3).

### **Figure 4.3**

*Actor meta-accuracy for cooperative behaviours across all motive and trait items*

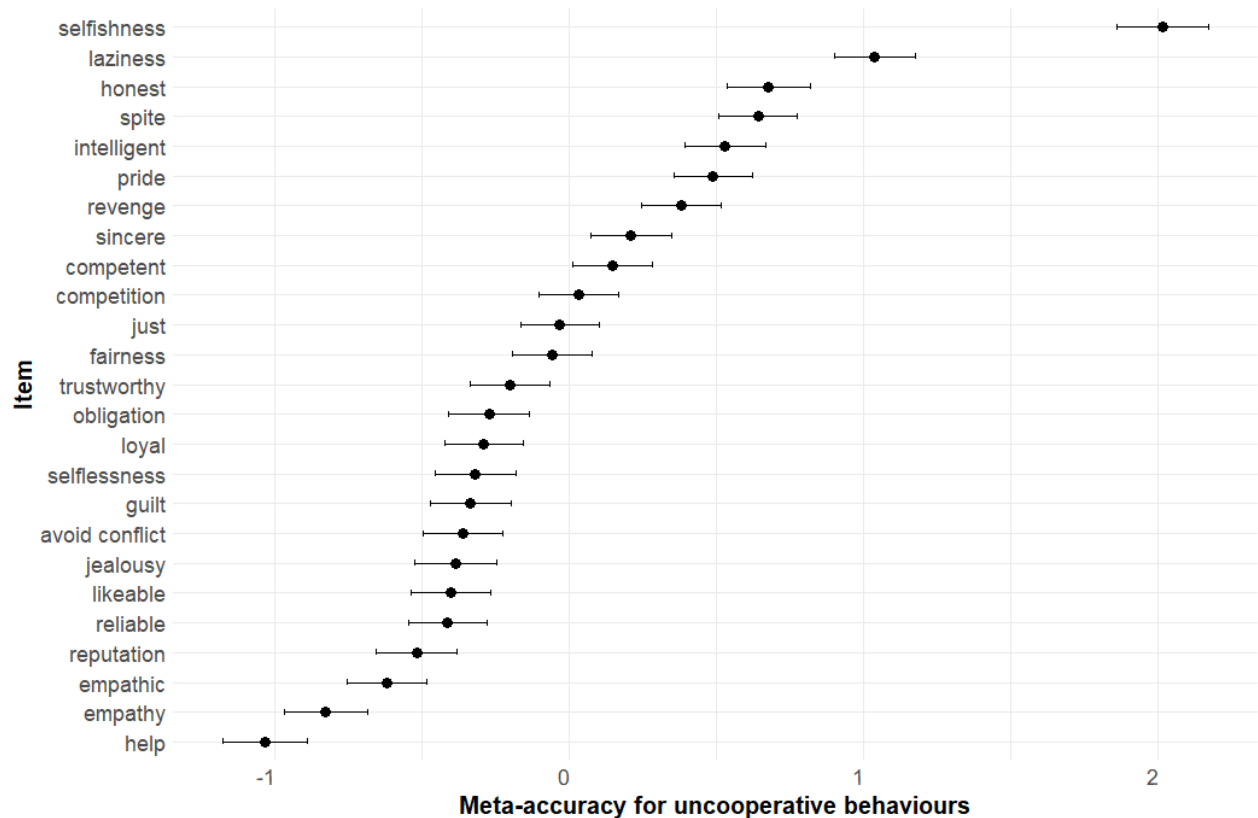


*Note.* Dots represent estimated marginal means (EMMs) and error bars signify 95% confidence intervals (CIs). Higher values indicate greater meta-accuracy, while lower values indicate lower meta-accuracy.

For uncooperative behaviours, actors exhibited the highest levels of meta-accuracy for the items selfishness, laziness, honesty, spite, and intelligence, while displaying the least accuracy for the items related to helping others, empathy, being empathic, reputation management, and reliability (Figure 4.4).

**Figure 4.4**

*Actor meta-accuracy for uncooperative behaviours across all motive and trait items*



*Note.* Dots represent EMMs and error bars signify 95% CI. Higher values indicate greater meta-accuracy, while lower values indicate lower meta-accuracy.

### 4.3.3 Meta-Bias

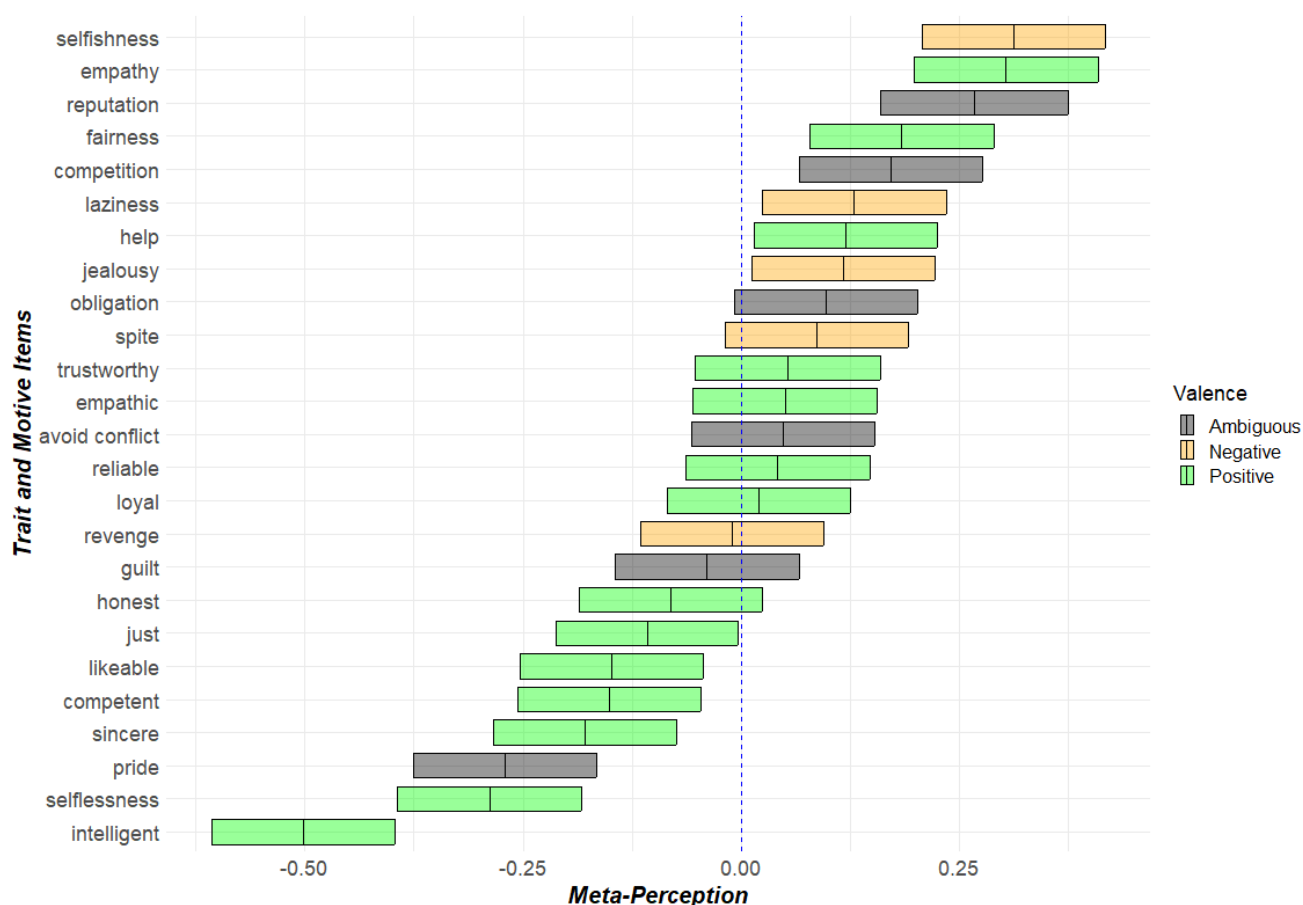
To examine systematic biases in meta-perception, I variable-centred the data and interacted the item variable with meta-accuracy (i.e., actor meta-perception minus the average observer evaluation per item) in the base meta-accuracy models before computing marginal means. Because the data is variable-centred, estimates which significantly differ from zero can be interpreted as systematic directional bias.

Actor participants did not exhibit a strong systematic bias when predicting how their cooperative behaviours would be evaluated (Figure 4.5). Actors anticipated being viewed as less intelligent, competent, likable, and just, as well as less motivated by selflessness and pride than they actually were. With the exception of pride, this underestimation pattern (as indicated by the estimates to the left side of the blue dashed line) suggests a negativity bias in the actors' expectations. However, while actors overestimated (as shown by the estimates to the right side of the blue dashed line) the more undesirable motive attributions of selfishness, reputation concerns, competition, laziness, and jealousy, they also overestimated the more desirable

motive attributions of empathy, fairness, and the desire to help others. Thus, the overall results indicate that actors did not exhibit a strong systematic bias in expecting more negative evaluations than they actually received for their cooperative behaviours.

**Figure 4.5**

*Visualisation of systematic biases in cooperative behaviour meta-perceptions using estimated marginal means.*



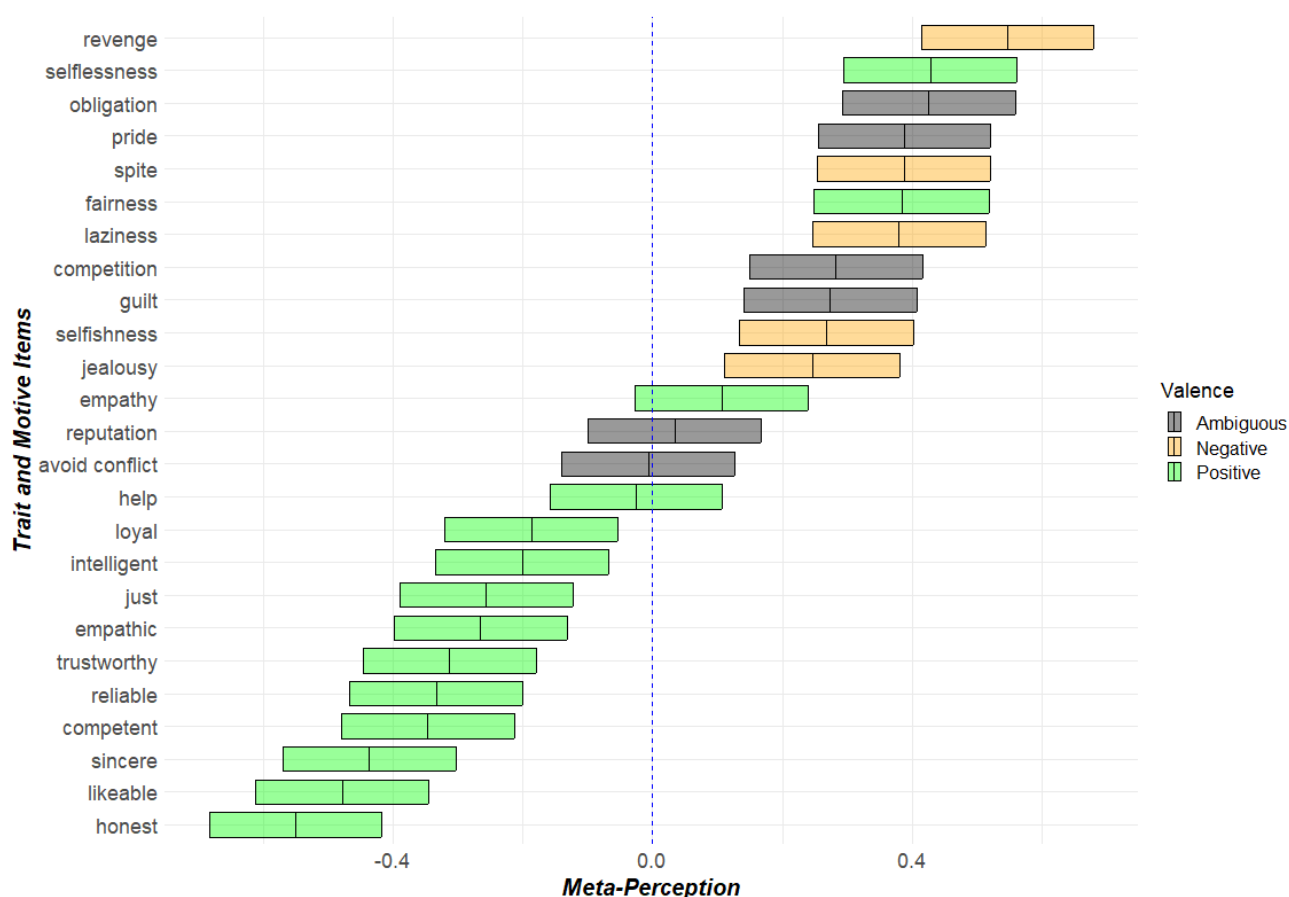
*Note.* This figure displays unstandardised estimated marginal means of actor meta-perceptions for various traits and motives, represented by the horizontal lines within each crossbar. The bars indicate the 95% CIs for these estimates. The dashed blue line at zero represents mean-level accuracy, meaning that estimates which deviate from zero can be interpreted as directional bias (i.e., items to the left represent underestimation whilst items to the right of the line represent overestimation).

Conversely, when predicting how they would be perceived for their uncooperative behaviours, actors exhibited a more systematic negativity bias, indicating that they tended to expect more

negative (or less positive) evaluations than they actually received (Figure 4.6). Specifically, actors overestimated how much observers would attribute their uncooperative behaviour to the negatively valenced motives revenge, obligation, pride, spite, laziness, competition, guilt, selfishness, and jealousy (but also selflessness and fairness). Additionally, they underestimated all (positively valenced) trait attributions—particularly honesty, likability, and sincerity—as well as the motive of helping others.

**Figure 4.6**

*Visualisation of systematic biases in uncooperative behaviour meta-perceptions using estimated marginal means.*



*Note.* This figure displays unstandardised estimated marginal means of actor meta-perceptions for various traits and motives, represented by the horizontal lines within each crossbar. The bars denote 95% confidence intervals. The dashed blue line at zero represents mean-level accuracy, meaning that estimates which deviate from zero can be interpreted as directional bias (i.e., items



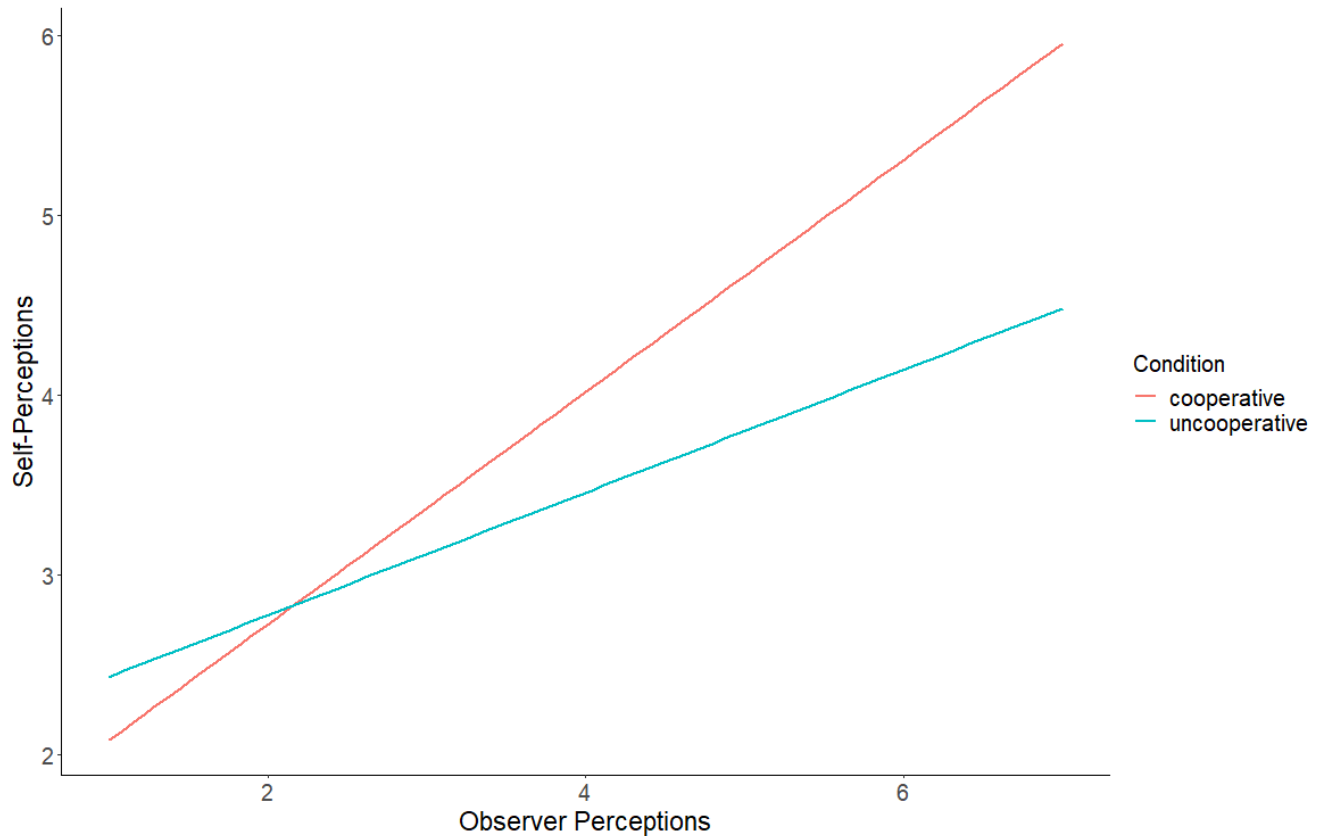
to the left represent underestimation whilst items to the right of the line represent overestimation).

#### 4.3.4 Observer Accuracy

In addition to examining actor meta-accuracy, this study also investigated observer accuracy. In simple terms, observer accuracy refers to the alignment between how observers perceive actors and the actors' actual motives and traits, as self-reported by the actors. Figure 4.7 visualises this alignment. Observers were able to accurately assess the self-reported motives and traits of actors for both cooperative ( $b = 0.71$ , 95% CI [0.68, 0.74],  $t(210.54) = 43.60$ ,  $p < .0001$ ) and uncooperative ( $b = 0.39$ , 95% CI [0.36, 0.42],  $t(176.44) = 22.39$ ,  $p < .0001$ ) behaviours. As observer accuracy may be influenced by how much observers rely on their preconceived notions or general knowledge about what motives and traits are common in the population, rather than only focusing on the specific motives and traits exhibited in the behaviour they are evaluating, caution is needed when interpreting baseline models.

#### Figure 4.7

*Relationship between observer-perceptions and actors' self-perceptions for cooperative and uncooperative behaviours*



*Note.* The figure presents the raw, uncentered data showing the relationship between observer-perceptions (how observers perceived actors) and corresponding actor self-perceptions (how actors viewed themselves) for both cooperative and uncooperative behaviours.

Distinctive accuracy models therefore controlled for the average distribution of motives and traits by examining how well estimated values aligned with actual values, while also assessing how closely estimated values matched the normative profile. For both behaviours, observers displayed distinctive accuracy (cooperative:  $b = 0.09$ , 95% CI [0.06, 0.12],  $t(171.93) = 6.67$ ,  $p < .0001$ ; uncooperative:  $b = 0.23$ , 95% CI [0.20, 0.27],  $t(179.35) = 13.27$ ,  $p < .0001$ ), as well as normative accuracy (cooperative:  $b = 0.87$ , 95% CI [0.82, 0.92],  $t(250.78) = 37.27$ ,  $p < .0001$ ; uncooperative:  $b = 0.76$ , 95% CI [0.70, 0.82],  $t(218.70) = 23.57$ ,  $p < .0001$ ). These results indicate that observers' perceptions aligned not only with the distinctiveness of the actor they were evaluating but also with the normative profile, which reflects the distribution of motive and trait means across the actor sample.

I also found evidence of observer insight, operationalised as the linear relationship between observer judgements and actors' self-perceptions while controlling for actors' meta-perceptions, for both cooperative ( $b = 0.39$ , 95% CI [0.34, 0.43],  $t(167.21) = 17.56$ ,  $p < .0001$ )

and uncooperative ( $b = 0.25$ , 95% CI [0.20, 0.27],  $t(177.03) = 12.74$ ,  $p < .0001$ ) behaviours. This means that observers perceived actors' true motives beyond how actors' thought they would be perceived.

I additionally found a statistically significant relationship between observers' attributions and actors' expectations for how they would be perceived (opaqueness) for cooperative ( $b = 0.40$ , 95% CI [0.36, 0.44],  $t(157.54) = 18.65$ ,  $p < .0001$ ) and uncooperative ( $b = 0.24$ , 95% CI [0.20, 0.28],  $t(165.30) = 12.76$ ,  $p < .0001$ ) actions. Opaueness is exhibited when observers' judgments are associated with actors' meta-perceptions above actors' self-reported actual motives and traits, meaning that observers perceived some items inaccurately and in a manner that actors anticipated. See Table 4.4 (cooperative behaviours) and Table 4.5 (uncooperative behaviours) for details on the three main models fit for each behaviour: baseline accuracy models, distinctive accuracy models, and insight models.

**Table 4.4**

*Observer Judgement Accuracy for Cooperative Behaviours Across Models*

| <i>Predictors</i> | <b>Base Accuracy Model</b> |          |                  | <b>Distinctive Accuracy Model</b> |          |                  | <b>Insight Model</b>   |          |                  |
|-------------------|----------------------------|----------|------------------|-----------------------------------|----------|------------------|------------------------|----------|------------------|
|                   | <i>b</i> (CI)              | <i>t</i> | <i>p</i>         | <i>b</i> (CI)                     | <i>t</i> | <i>p</i>         | <i>b</i> (CI)          | <i>t</i> | <i>p</i>         |
| (Intercept)       | 0.04<br>(-0.05 – 0.14)     | 0.90     | .37              | 0.04<br>(-0.05 – 0.14)            | 0.89     | .37              | 0.04<br>(-0.08 – 0.17) | 0.68     | 0.50             |
| Self-Perceptions  | 0.71<br>(0.68 – 0.74)      | 43.60    | <b>&lt; .001</b> | 0.09<br>(0.06 – 0.12)             | 6.67     | <b>&lt; .001</b> | 0.39<br>(0.34 – 0.43)  | 17.56    | <b>&lt; .001</b> |
| Normative Profile |                            |          |                  | 0.87<br>(0.82 – 0.92)             | 37.27    | <b>&lt; .001</b> |                        |          |                  |
| Meta-Perceptions  |                            |          |                  |                                   |          |                  | 0.40<br>(0.36 – 0.44)  | 18.65    | <b>&lt; .001</b> |

*Note.* This table presents observer judgement accuracy for cooperative behaviours across the three main model types fit in this chapter. Estimates (*b*), 95% confidence intervals (CIs), *t*-values, and *p*-values are reported. Significant results are bolded. The **baseline accuracy model**

tested for a linear relationship between observer perceptions and actor self-perceptions. As observer accuracy depends both on an understanding of an actors' unique traits and motives relative to other actors (*distinctive meta-accuracy*) and an understanding of which motives and traits are common in the population (*normative accuracy*), the distinctive accuracy model tested for a linear relationship between observer-perceptions and the normative profile (the grand mean centred distribution of the self-perception means across actors) on top of the baseline models' linear relationship between observer-perceptions and actor self-perceptions. Here, the value for self-perceptions reflects distinctive accuracy, while the value for the normative profile reflects normative accuracy. The *insight model* tested for a linear relationship between observer-perceptions and actor self-perceptions (*insight*) while controlling for actors' meta-perceptions (*opaqueness*). Observer insight reflects the observer's ability to accurately perceive the disparity between an actor's actual motives and traits and the actor's predicted perceptions. Conversely, opaqueness is exhibited when observers perceptions are associated with actors' meta-perceptions above actors' self-reported actual motives and traits, which would mean that observers perceive some items inaccurately and in a manner that actors anticipated. Additionally, as I used person centring (cluster centring within actor) so that meta- and observer-perceptions were anchored to the mean validation criterion within each actor (i.e., meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor), a value of zero represents the average point estimate accuracy within actor, and deviations from zero provide insights into average point-estimate inaccuracies.

**Table 4.5**

*Observer Judgement Accuracy for Uncooperative Behaviours Across Models*

| <i>Predictors</i>    | <b>Base Accuracy Model</b>  |           |                  | <b>Distinctive Accuracy Model</b> |           |                  | <b>Insight Model</b>        |           |                  |
|----------------------|-----------------------------|-----------|------------------|-----------------------------------|-----------|------------------|-----------------------------|-----------|------------------|
|                      | <i>b (CI)</i>               | <i>t</i>  | <i>p</i>         | <i>b (CI)</i>                     | <i>t</i>  | <i>p</i>         | <i>b (CI)</i>               | <i>t</i>  | <i>p</i>         |
| (Intercept)          | -0.18<br>(-0.31 --<br>0.06) | -2.87     | <b>.004</b>      | -0.18<br>(-0.31 --<br>0.06)       | -2.87     | <b>.005</b>      | -0.20<br>(-0.34 --<br>0.05) | -2.73     | <b>.007</b>      |
| Self-<br>Perceptions | 0.39<br>(0.36 -- 0.<br>42)  | 22.3<br>9 | <b>&lt; .001</b> | 0.23<br>(0.20 -- 0.27<br>)        | 13.2<br>7 | <b>&lt; .001</b> | 0.24<br>(0.20 -- 0.27<br>)  | 12.7<br>4 | <b>&lt; .001</b> |

|                      |                           |                           |                     |
|----------------------|---------------------------|---------------------------|---------------------|
| Normative<br>Profile | 0.76<br>(0.70 – 0.82<br>) | 23.5<br>7                 | < .001              |
| Meta-<br>Perceptions |                           | 0.24<br>(0.20 – 0.28<br>) | 12.7<br>6<br>< .001 |

*Note.* This table presents observer judgement accuracy for uncooperative behaviours across the three main model types fit in this chapter. Estimates (*b*), 95% confidence intervals (CIs), *t*-values, and *p*-values are reported. Significant results are bolded. The ***baseline accuracy model*** tested for a linear relationship between observer perceptions and actor self-perceptions. As observer accuracy depends both on an understanding of an actors' unique traits and motives relative to other actors (*distinctive meta-accuracy*) and an understanding of which motives and traits are common in the population (*normative accuracy*), the distinctive accuracy model tested for a linear relationship between observer-perceptions and the normative profile (the grand mean centred distribution of the self-perception means across actors) on top of the baseline models' linear relationship between observer-perceptions and actor self-perceptions. Here, the value for self-perceptions reflects distinctive accuracy, while the value for the normative profile reflects normative accuracy. The ***insight model*** tested for a linear relationship between observer-perceptions and actor self-perceptions (*insight*) while controlling for actors' meta-perceptions (*opaqueness*). Observer insight reflects the observer's ability to accurately perceive the disparity between an actor's actual motives and traits and the actor's predicted perceptions. Conversely, opaqueness is exhibited when observers perceptions are associated with actors' meta-perceptions above actors' self-reported actual motives and traits, which would mean that observers perceive some items inaccurately and in a manner that actors anticipated. Additionally, as I used person centring (cluster centring within actor) so that meta- and observer-perceptions were anchored to the mean validation criterion within each actor (i.e., meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor), a value of zero represents the average point estimate accuracy within actor, and deviations from zero provide insights into average point-estimate inaccuracies.

When behaviour condition was inserted as moderator of observer accuracy, it significantly influenced both distinctive judgement accuracy ( $b = 0.12$ , 95% CI [0.11, 0.14],  $t(78942.28) =$

16.26,  $p < .0001$ ) and normative accuracy ( $b = -0.09$ , 95% CI [-0.11, -0.06],  $t(82127.48) = -7.07$ ,  $p < .0001$ ). Specifically, observers displayed higher distinctive accuracy for uncooperative compared to cooperative behaviours, but higher normative accuracy for cooperative behaviours.

### 4.3.5 Secondary Analyses

Both actors ( $b = 0.52$ , 95% CI [0.23, 0.80],  $t(158) = 3.61$ ,  $p = .0004$ ) and observers ( $b = 0.83$ , 95% CI [0.70, 0.95],  $t(424) = 12.87$ ,  $p < .0001$ ) correctly anticipated that actor meta-perceptions and observer evaluations would align more for cooperative than for uncooperative behaviours. Actors ( $b = 1.23$ , 95% CI [1.00, 1.46],  $t(158) = 10.75$ ,  $p < .0001$ ) and observers ( $b = 0.17$ , 95% CI [0.02, 0.31],  $t(424) = 2.27$ ,  $p = .02$ ) also believed that cooperative behaviours were more indicative of actor character than uncooperative behaviours were, although this view was more pronounced for actor participants.

Lastly, to assess which items were rated as more applicable to either cooperative or uncooperative behaviours, I conducted mixed-effects ANOVAs (separately for self-perceptions, meta-perceptions and observer judgements) and performed pairwise comparisons on the estimated marginal means for the interaction between item and condition, while adjusting for multiple comparisons with the Bonferroni correction (Table 4.6). There were significant main effects of item and condition, as well as their interaction for all three rating types (all  $ps < .0001$ ).

**Table 4.6**

*Estimates (b), t-values, and p-values for the applicability of motives and traits to cooperative versus uncooperative behaviours*

| Motive/Trait Items | Observer Perceptions (O) |          |          | Meta-Perceptions (A) |          |          | Self-Perceptions (A) |          |          |
|--------------------|--------------------------|----------|----------|----------------------|----------|----------|----------------------|----------|----------|
|                    | <i>b</i>                 | <i>t</i> | <i>p</i> | <i>b</i>             | <i>t</i> | <i>p</i> | <i>b</i>             | <i>t</i> | <i>p</i> |
| Reputation         | 2.38                     | 40.88    | <.0001   | 2.60                 | 31.84    | <.0001   | 1.59                 | 19.38    | <.0001   |
| Laziness           | -2.66                    | -45.76   | <.0001   | -3.08                | -37.79   | <.0001   | -2.46                | -30.00   | <.0001   |
| Pride              | 0.08                     | 1.41     | 0.16     | -0.58                | -7.11    | <.0001   | -0.35                | -4.25    | <.0001   |
| Spite              | -1.85                    | -31.75   | <.0001   | -2.24                | -27.48   | <.0001   | -1.80                | -21.90   | <.0001   |
| Jealousy           | -0.77                    | -13.33   | <.0001   | -0.98                | -12.04   | <.0001   | -0.58                | -7.04    | <.0001   |
| Competition        | -0.51                    | -8.70    | <.0001   | -0.68                | -8.27    | <.0001   | -0.56                | -6.80    | <.0001   |
| Revenge            | -1.37                    | -23.60   | <.0001   | -1.98                | -24.30   | <.0001   | -1.43                | -17.40   | <.0001   |
| Selflessness       | 2.99                     | 51.55    | <.0001   | 2.31                 | 28.33    | <.0001   | 2.10                 | 25.58    | <.0001   |
| Obligation         | 1.81                     | 31.11    | <.0001   | 1.53                 | 18.70    | <.0001   | 1.07                 | 13.01    | <.0001   |
| Avoiding conflict  | 0.42                     | 7.15     | <.0001   | 0.49                 | 6.06     | <.0001   | 0.09                 | 1.10     | 0.27     |
| Helping others     | 4.32                     | 74.38    | <.0001   | 4.43                 | 54.26    | <.0001   | 4.21                 | 51.34    | <.0001   |
| Fairness           | 2.20                     | 37.92    | <.0001   | 2.09                 | 25.66    | <.0001   | 1.84                 | 22.47    | <.0001   |
| Empathy            | 3.47                     | 59.72    | <.0001   | 3.62                 | 44.41    | <.0001   | 3.45                 | 42.01    | <.0001   |
| Guilt              | 0.20                     | 3.41     | .0007    | -0.08                | -0.97    | 0.33     | -0.45                | -5.52    | <.0001   |
| Selfishness        | -2.71                    | -46.66   | <.0001   | -2.74                | -33.54   | <.0001   | -2.40                | -29.27   | <.0001   |
| Trustworthy        | 2.63                     | 45.22    | <.0001   | 3.01                 | 36.85    | <.0001   | 2.53                 | 30.78    | <.0001   |
| Intelligent        | 1.67                     | 28.77    | <.0001   | 1.41                 | 17.26    | <.0001   | 0.99                 | 12.06    | <.0001   |
| Sincere            | 2.14                     | 36.83    | <.0001   | 2.39                 | 29.25    | <.0001   | 2.05                 | 24.97    | <.0001   |
| Honest             | 1.42                     | 24.40    | <.0001   | 1.87                 | 22.96    | <.0001   | 1.45                 | 17.66    | <.0001   |
| Loyal              | 2.76                     | 47.52    | <.0001   | 2.98                 | 36.49    | <.0001   | 2.52                 | 30.72    | <.0001   |
| Just               | 2.15                     | 37.06    | <.0001   | 2.32                 | 28.45    | <.0001   | 1.79                 | 21.81    | <.0001   |
| Reliable           | 3.04                     | 52.28    | <.0001   | 3.42                 | 41.90    | <.0001   | 2.94                 | 35.86    | <.0001   |
| Empathic           | 3.30                     | 56.76    | <.0001   | 3.61                 | 44.25    | <.0001   | 3.04                 | 37.08    | <.0001   |
| Likeable           | 2.91                     | 50.09    | <.0001   | 3.23                 | 39.57    | <.0001   | 2.96                 | 36.05    | <.0001   |
| Competent          | 2.08                     | 35.86    | <.0001   | 2.29                 | 28.11    | <.0001   | 1.75                 | 21.26    | <.0001   |

*Note.* This table presents the results of pairwise comparisons on the estimated marginal means for the interaction between item and condition, conducted separately for observer, meta- and self-perceptions. A Bonferroni correction was applied to adjust for multiple comparisons across all rating types, with an adjusted significance level of  $\alpha = .00067$  (i.e.,  $.05/75$ ). "O" stands for perceptions provided by observers, while "A" stands for perceptions provided by actors. For easy viewing, red indicates the item is significantly more applicable to cooperative behaviours, blue indicates the item is significantly more applicable to uncooperative behaviours, and grey indicates no significant difference.

Observers perceived the motives of helping others, empathy, selflessness, reputation management, fairness, obligation, and avoiding conflict, as well as all traits (particularly empathic, reliable, likeable, loyal, and trustworthy), as more applicable to cooperative behaviours compared to uncooperative behaviours. In contrast, the motives selfishness, laziness, spite, revenge, jealousy, and competition were seen as more applicable to uncooperative behaviours. Pride and guilt were not distinctly associated with either cooperative or uncooperative behaviours.

Actors correctly anticipated these views, with meta-perceptions largely mirroring observer perceptions, except for pride, which actors expected to be significantly more associated with uncooperative behaviours. Actors' self-perceptions followed a similar pattern, but they indicated being driven significantly more by guilt during uncooperative actions compared to cooperative ones. Additionally, there was no significant difference in how much they reported being motivated by conflict avoidance when engaging in cooperative versus uncooperative behaviours.

### *Correlations Between Trait and Motive Items in Cooperative and Uncooperative Behaviours*

In addition to identifying which traits and motives were more strongly associated with cooperative versus uncooperative behaviours, the relationships among these traits and motives within each behavioural context were examined. These correlations reveal how closely aligned different traits and motives were.

The correlations among personality traits for cooperative behaviours ranged from  $r = 0.03$  (between empathic and competent) to  $r = 0.66$  (between honest and trustworthy). For uncooperative behaviours, the correlations ranged from  $r = 0.39$  (between intelligent and loyal) to  $r = 0.78$  (between reliable and trustworthy). The correlations among motives for cooperative behaviours ranged from  $r = 0.00$  (between revenge and reputation, and between competition and empathy) to  $r = 0.61$  (between spite and jealousy). For uncooperative behaviours, the correlations ranged from  $r = 0.00$  (between spite and conflict avoidance, and between spite and helping another) to  $r = 0.78$  (between revenge and spite).

When examining the correlation between motives and traits, for cooperative behaviours, the correlations ranged from  $r = 0.00$  (between the motive conflict avoidance and the trait sincere, between the motive pride and the traits loyal and just, as well as between the motive competition and the trait reliable) to  $r = 0.64$  (between the motive empathy and the trait empathic). For uncooperative behaviours, correlations ranged from  $r = 0.00$  (between the trait competent and the motives spite and revenge) to  $r = 0.59$  (between the motive fairness and the trait just).

Since responses across the different rating types (meta-perceptions, self-perceptions, and observer perceptions) were highly similar, I reported only the correlations between motive and trait items for meta-perceptions of cooperative and uncooperative behaviours. This approach helps avoid redundancy and prevents the thesis from being overloaded with repetitive tables, thereby enhancing the accessibility of findings. I chose to report correlations for meta-



perceptions because they are the primary perspective of interest in this chapter. Given the high consistency between the rating types, as demonstrated throughout the results section of this chapter, reporting the correlation matrices for meta-perceptions alone should sufficiently illustrate the relationships between items. See Appendix C4 for the correlation matrices.

## 4.4 Discussion

For the present study I solicited written descriptions of recent cooperative and uncooperative behaviours with accompanying self- and meta-perceptions. A separate set of participants then judged the authors based on their behaviours. This method allowed me to directly test participants' meta-perceptive accuracy, as well as the extent to which observers could predict the self-reported motives and traits of actors in the context of real-world cooperative and uncooperative behaviours. In addition to focusing on stable constructs, as has commonly been the focus, this study aimed to ascertain whether individuals are aware of the motives they are attributed and how others perceive them based on specific actions. By investigating whether we know how our cooperative and uncooperative actions influence motive attributions and our moral reputation, I aimed to further our understanding on reputation and reputation management in everyday life.

I predicted that actors would exhibit meta-perceptive accuracy for both cooperative and uncooperative behaviours, but that cooperative behaviours would be linked to higher levels of meta-perspective accuracy than uncooperative behaviours. These predictions were largely supported. Actors displayed baseline meta-accuracy (alignment between observer- and meta-perceptions), distinctive accuracy (knowledge of how they were uniquely perceived relative to the average actor; Biesanz, 2010; Furr, 2008), and normative accuracy (reflecting shared understandings of trait and motive perceptions across the actor sample; Wood & Furr, 2016). Additionally, they showed insight into how they would be misperceived by observers (Carlson, 2016) for both cooperative and uncooperative behaviours.

In line with predictions, both actors and observers correctly anticipated that actor meta-perceptions and observer judgements would align more closely for cooperative compared to uncooperative actions. However, while the profile agreement between actors' expected evaluations (meta-perceptions) and actual evaluations (observer judgments) was stronger for cooperative than uncooperative behaviours, the componential approach of the SAM revealed that this was primarily driven by higher levels of normative accuracy for cooperative behaviours. Distinctive accuracy was higher for uncooperative behaviours than for cooperative ones. Thus, while actors had a clearer understanding of how the average person engaging in cooperative actions would be perceived, they demonstrated greater awareness of how they would be uniquely perceived for uncooperative actions. One possible explanation for this finding is that cooperative behaviours tend to align more closely with societal norms. This normativity may make cooperative actions less likely to be seen as uniquely revealing about

the actor, potentially leading them to be evaluated in a more generalised manner. As a result, actors might find it harder to distinguish how their specific cooperative actions are uniquely perceived. In contrast, uncooperative behaviours may stand out more because they deviate more from social norms and can carry higher social costs (e.g., damage to relationships or harm to reputation). Actors might, therefore, pay closer attention to how their uncooperative actions are perceived. Such increased attention could contribute to a greater understanding of how these behaviours uniquely shape others' impressions of them. Additionally, while cooperative behaviours may elicit more frequent feedback (Gallrein et al., 2019), the feedback actors receive for uncooperative behaviours might be more distinctive when it occurs. Feedback on uncooperative actions may be more negative and, therefore, could carry more weight in memory and reflection than the more positive feedback cooperative behaviours may elicit (Baumeister et al., 2001).

At the same time, actors displayed a transparency bias for both behaviour conditions by believing their motives would be perceived more accurately by observers than they were. In line with this, actors' meta- and self-perceptions were highly correlated for both cooperative and uncooperative behaviours, which suggests that actors did not expect observers' perceptions to significantly deviate from how they perceive themselves. This is further evidence of a transparency bias (Gilovich et al., 1998), where actors believe their feelings or thoughts are more transparent to observers than they actually are, resulting in overestimations of how accurately observers will perceive them. My chosen methods could have somewhat enhanced this transparency bias, as asking participants to write about their recent actions might have enabled actors to include explanations for their behaviour, which they may not generally voice in face-to-face interactions.

I furthermore expected meta-accuracy to vary based on item. As meta-accuracy tends to be higher for observable traits (items linked to more external cues or behaviours) as opposed to those that are more internal or evaluative (Carlson & Kenny, 2012; Stopfer et al., 2014), I expected actors to exhibit higher levels of meta-perceptive accuracy for more observable items such as cooperativeness compared to more internal motivation items. Furthermore, since accuracy increases when cues are relevant to the traits being assessed (Hirschmüller et al., 2015; Letzring & Funder, 2018), I anticipated that meta-accuracy would be higher for more relevant items. Item was a significant moderator of meta-accuracy for both cooperative and uncooperative behaviours. For cooperative behaviours, actors were most accurate in predicting the traits reliability, trustworthiness, likeability, and honesty, as well as the motive of helping

others. They were least accurate in predicting the motive items revenge, laziness, spite, jealousy, competition, and guilt. This suggests that actors were generally more accurate at predicting evaluations for relevant trait attributions while struggling to predict motive items that do not strongly drive cooperative actions. Thus, meta-accuracy was lower for internal motivations than for (possibly) more observable traits, though traits themselves are likely only more visible when directly related to the specific action under review. For uncooperative behaviours, actors exhibited the highest levels of meta-accuracy for the motives selfishness, laziness, and spite, as well as the traits honesty and intelligence. They displayed the least accuracy for the motives helping others, empathy, and reputation management, as well as the traits being empathic, reliable, and likeable. Thus, actors were more accurate in predicting motives and traits plausibly linked to uncooperative actions. This highlights that accuracy in predicting motives and traits is influenced by the relevance and, consequently, the visibility of the items being assessed. Relevance likely played a larger role than visibility in this context, as items were only more visible when they were more relevant. However, this is likely due to the limited visibility of the items used in this study, as they were primarily internal.

Social subjective status, reputation concern, and social perspective taking did not moderate meta-accuracy. Future studies ought to additionally examine moderators of observer accuracy, as a prior study found multiple moderators of observer accuracy which were distinct from moderators affecting meta-accuracy (Lees et al., 2022).

Individuals, even when displaying meta-accuracy, may still err by systematically over- or underestimating how favourably others perceive them. In situations where errors have varying cost magnitudes, individuals should favour strategies which minimize the more expensive error type by increasing the occurrence of the cheaper type (Efferson et al., 2020; Kiyonari et al., 2000; McKay & Efferson, 2010; Yamagishi et al., 2007). Therefore, I hypothesised that actors would systematically underestimate how positively others view them. A negativity bias (expecting others to view them less favourably than they actually do) might keep individuals driven to monitor and control their conduct and reputation. Not doing so when one ought to can bring serious consequences, such as permanent harm to relationships or social isolation (FeldmanHall & Shenhav, 2019). In line with predictions, when considering their recent uncooperative actions, actors exhibited a systematic negative bias. While, overall, actors had a slight tendency towards a negativity bias, they did not, however, exhibit a strong systematic bias for cooperative behaviours. It has been proposed (Welker et al., 2023) that the systematic negative meta-bias found in recent research (Boothby et al., 2018; Elsaadawy & Carlson, 2022;

Mastroianni et al., 2021; Wolf et al., 2021) is specific to conversations. The present study shows that this is not the case, but the reasoning behind this idea may shed light on why actors systematically underestimated how positively they would be regarded by others when the focus was on their uncooperative behaviours. During conversations, individuals may over-focus on social blunders and cues of rejection (Gilovich et al., 2000). Searching for negative evaluations during these interactions may make negative thoughts more salient, leading to pessimistic meta-perceptions (Boothby et al., 2018). Similarly, when engaging in uncooperative behaviours (perhaps more so than cooperative behaviours), individuals might be more focused on their reputation and more attuned to potential negative evaluations, which can result in a negativity bias.

Similar to actors, observers were also expected to provide accurate motive and trait attributions. Results were in line with predictions. When evaluating cooperative and uncooperative behaviours, observers exhibited baseline, distinctive, and normative accuracy. Furthermore, observers exhibited insight in both behaviour conditions, meaning that observers perceived actors' true motives beyond how actors' thought they would be perceived. At the same time, I found evidence of opaqueness, which means that observers perceived some items inaccurately and in a manner that actors anticipated (this occurs when observers' judgments are associated with actors' meta-perceptions above actors' self-reported actual motives and traits). Overall, observers, just like actors, displayed significantly higher profile agreement and normative accuracy for cooperative compared to uncooperative behaviours, and higher distinctive accuracy for uncooperative behaviours.

When examining which items were rated as more applicable to cooperative or uncooperative behaviours, participants provided responses as expected. Observers viewed positively valenced items (e.g. helping others, empathy, selflessness) as well as positive personality traits (e.g. empathic, reliable, likeable) as more applicable to cooperative behaviour descriptions, whilst judging negatively valenced motives (e.g. selfishness, laziness, spite) as more applicable to uncooperative behaviours. Actors anticipated these views, with meta-perceptions largely mirroring observer perceptions. Actors' self-perceptions followed a similar pattern.

Contrary to predictions and prior literature (Reeder & Brewer, 1979), uncooperative behaviours were not perceived as more indicative of actor character. For actors this finding is plausible, as people are motivated to perceive themselves as moral (Fiske, 2014; Leary, 2007), which is likely easier when considering the cooperative action as more key to, or in line with, ones character than the uncooperative action. Moreover, the uncooperative behaviours we normally

hear about may be more conspicuous or bigger incidents than those reported in this study, which might be more mundane, commonplace, and not always directly observed in daily life. Consequently, observers too may be disinclined to attribute greater significance to such behaviours as markers of character, potentially due to their own occasional lapses in cooperative conduct, such as sporadic neglect of household chores, reluctance to assist colleagues, or feigning busyness to warrant respite. Uncooperative behaviours are perhaps only perceived as particularly informative regarding someone's character when they are extreme or uncommon (Ditto & Jemmott, 1989; Fiske, 1980; Kelley, 1967; McKenzie & Mikkelsen, 2007). Although participants provided realistic accounts of uncooperative behaviours, it is possible that they avoided sharing particularly severe, unusual, or socially undesirable actions, instead being more likely to describe common or mild uncooperative actions.

In general, participants' self-reports may have involved a form of reputation management, where they refrained from disclosing severe uncooperative actions to avoid negative perceptions or selectively chose to report cooperative behaviours that cast them in a positive light. They may also have adjusted the descriptions of their behaviours (e.g. including mitigating explanations for uncooperative actions or emphasising personal sacrifice for cooperative ones). Indeed, many participants included justifications for their actions, likely to explain their motives, reduce potential negative judgments, or highlight the value of their cooperative behaviours. This may have introduced another potential bias, as perceptions and their accuracy could be influenced by the narrative framing provided. Participants who justified their actions may have been viewed more favourably than those who described their actions more factually, without explanations. Those who included more explanations might also have been better at predicting others' judgments of them and may have been judged more accurately themselves (though these possibilities were not tested in this study). See Appendix C5 for a sample of participant responses, providing insight into the range of cooperative and uncooperative behaviours reported. It should also be noted that results might not generalise to the targets of uncooperative and cooperative behaviours. Observer participants were unaffected, neutral strangers, and actors were asked for meta-perceptions of the average person, not anyone involved in their stories.

Altogether, this approach produced a large and diverse set of naturalistic instances of cooperative and uncooperative behaviours, which allows for more ecologically valid inferences than a reliance on highly stylised and fictional scenarios could have yielded. The findings revealed nuanced patterns: while actors and observers demonstrated perceptive accuracy across

behaviours, they consistently displayed higher distinctive accuracy for uncooperative behaviours, whilst tracking more with the normative profile for cooperative actions. Moreover, actors exhibited a transparency bias, believing their motives were more apparent to others than they actually were. Notably, actors also exhibited a systematic negative bias in how they perceived others' evaluations of their uncooperative behaviours, suggesting a tendency to underestimate positive perceptions. Overall, this study sheds light on the complex interplay between behaviours and perceptions in everyday social interactions. Future research should continue to explore these dynamics, considering other contextual factors that may influence meta-perceptive accuracy and bias. Understanding these processes is crucial not only for advancing psychological theory but also for informing strategies for effective communication, interpersonal relations, and reputation management.

# **Chapter 5**

## **Understanding Punisher Motives and Reputation: Meta-Perceptions and Observer Inferences Surrounding Punishment**



# **5.1 Investigating Observer Perceptions of Motives in Second- and Third-Party Punishers through Vignettes**

## **5.1.1 Introduction**

To reiterate, punishment, in this thesis defined as a response to an offence via inflicting some costs on the transgressor (Boyd & Richerson, 1992; Clutton-Brock & Parker, 1995), is a mechanism with which cooperation can be promoted and norm violations can be deterred (Balliet et al., 2011; Boyd & Richerson, 1992; Gintis et al., 2008). However, punishers are not consistently cooperative, punishment does not always promote cooperation or improve group outcomes, and is often motivated by payoff disparities rather than solely by the act of cheating or experiencing losses (Raihani & Bshary, 2019). Wu et al. (2022) and Molho and Wu (2021) summarise that the underlying factors contributing to the positive selection of punishment also encompass the potential for the punisher to improve their own resources or social standing relative to the target (Deutchman et al., 2021; Raihani & Bshary, 2019) as well as opportunities for reputational benefits (Mifune et al., 2020; Raihani & Bshary, 2015b).

As emphasised in earlier chapters, evidence on the reputational consequences of punishment is mixed (Raihani & Bshary, 2015b). Even punishers themselves both advertise (Kurzman et al., 2007; Piazza & Bering, 2008) and hide (Rockenbach & Milinski, 2011) their punitive behaviour, indicating that punishment sometimes does and sometimes does not yield reputation benefits. Central to several chapters of this thesis is the suggestion that punitive actions can affect reputation by either signalling that the punisher is formidable or alternatively, if self-serving motives are unlikely, can signal the punisher's cooperative intent (Raihani & Bshary, 2015c). People engage in punishment both when they themselves were the victim of an offense (second-party punishment; SPP) (Fehr & Gächter, 2002; Henrich et al., 2006) and when they are observing the offence as a third-party (third-party punishment; TPP) (Fehr & Fischbacher, 2003; Marlowe et al., 2008, 2011). When comparing different game types, punishment is commonly evaluated negatively in public goods games with punishment (where the punisher is affected by free riders), but it is typically viewed positively in TPP games, where the punisher is not directly affected (compared in Li & Mifune, 2022). Vignette and economic games studies

have shown that third-party punishers are perceived as more trustworthy and less selfish (de Kwaadsteniet et al., 2019; Dhaliwal et al., 2020; Jordan et al., 2022). Yet, other studies indicate that third-party punishers can be evaluated as less trustworthy and more selfish than non-punishers (Eriksson et al., 2016; Heffner & FeldmanHall, 2019; Rai, 2022; Strimling & Eriksson, 2014). These complex findings could be explained by the motives attributed to punishment. Generally, observers are more likely to attribute ulterior motives to second-party than third-party punishers (Muñoz-Herrera & Nikiforakis, 2023).

As mentioned in the general introduction, Brañas-Garza et al. (2014) suggest that motives can be prosocial, selfish, and/or spiteful, with varying levels of ambiguity. Thus, even third-party punishers can be thought to be acting out of jealousy or competitiveness in certain contexts (e.g. Pedersen et al., 2013). For instance, punishers who are willing to pay high costs to punish are preferred over punishers who pay low costs, as selfish motives appear less likely (Nelissen, 2018). Owing to punishment being associated with multiple strategic goals, various additional motives reportedly underly punitive actions (Raihani & Bshary, 2019). For example, punishment has been linked to a desire for revenge (Bone & Raihani, 2015; Deutchman et al., 2021; Twardawski et al., 2022), status concerns (Redhead et al., 2021; Yamagishi et al., 2012), aversion to (disadvantageous) inequality (Bone et al., 2016; Bone & Raihani, 2015; Hechler & Kessler, 2022; Raihani & McAuliffe, 2012), and serves as a form of communication to, for instance, convey disapproval (Aharoni et al., 2022; Cushman et al., 2019; Funk et al., 2014; Sarin et al., 2021). Two broad functional explanations for punishment are deterrence and competition. According to the deterrence theory, punishment functions to discourage the violator (i.e., special deterrence) or others more broadly (i.e., general deterrence) from harming the punisher in future (Delton & Krasnow, 2017; McCullough et al., 2013). Motives of revenge or retribution are in line with this theory (Delton & Krasnow, 2017). Competitive punishment on the other hand functions to increase the punishers relative payoffs and fitness levels in comparison to transgressors (Bone & Raihani, 2015; Price et al., 2002; Raihani & Bshary, 2019). Motivations of inequity aversion are in line with this theory (Price et al., 2002). These functions are not mutually exclusive, and a combination thereof likely forms punishment (Deutchman et al., 2021).

Beyond investigating the motives underlying punitive behaviour, this study examined a lesser-explored aspect of punishment—the motives that observers attribute to punishers. In a recent study Muñoz-Herrera and Nikiforakis (2023) argued that observers are more likely to attribute ulterior motives to second-party rather than third-party punishers. This was largely based on

taking games with punishment in which observers were directly asked whether they thought the second-party or third-party punishing action was more biased. In another recent study using vignettes, Marshall and colleagues (2022) compared children's and adults' endorsement of three deontological and three consequentialist motives of third-party punishment. Deontological motives relate to notions of fairness and justice (e.g., 'just deserts'), whereas consequentialist motives are intended to bring about positive future consequences (e.g. deterrence). Although both endorsed consequentialist motives, children rejected the idea that people punish out of revenge, whilst adults considered 'just deserts' a probable motive for peer punishment.

Pursuing alternative methods to Marshall et al., I investigated the motives observers attribute to both second- and third-party punishers. Rather than focusing on a limited set of consequentialist and deontological motives, participants were presented with a more comprehensive array of 18 motive items spanning various functions (e.g. deterrence, reputational benefits, retribution, communication). Importantly, as they were working with children, Marshall et al. were limited to soliciting binary "yes" or "no" responses to motive items, which precludes the assessment of the strength of endorsement, a crucial aspect considering the complexity of punishing motives. Similarly, by focusing on children the exploration of adult peer punishment was neglected despite its common occurrence and inherent interest.

As the factors influencing how individuals perceive punishers and their motives is complex and remains uncertain, the present study aimed to examine attributions of punishing motives using vignettes, specifically focusing on the context of second-party punishment and third-party punishment. Participants read seven vignettes, each illustrating a distinct scenario involving a wrongdoing and subsequent punishment. Crucially, each vignette was randomly presented in one of two versions: either depicting second-party punishment or portraying third-party punishment. Field research suggests that individuals employ a range of punishing tactics, varying in cost and level of directness, to address norm violations (Balafoutas et al., 2014, 2016; Molho et al., 2020). Punishing tactics observed include physical and verbal confrontation, social exclusion, and reputation manipulation through gossip (Boehm, 1993; Cushman et al., 2019; Raihani & Bshary, 2019). Ethnographies even suggest punitive sentiment is often expressed through criticism or mocking of the target (Baumard 2010; Boehm 2011; Wiessner 2005). Consequently, the vignettes in this study included a broad range of direct and indirect punishing responses to various norm violations to enhance the ecological validity of findings,

especially as punishing methods beyond financial punishment are often overlooked (Molho et al., 2020; Molho & Wu, 2021).

I formed three main hypotheses. Firstly, I hypothesised that punishing actions would generally elicit positive perceptions of punishers. Although many studies indicate that punishers are disapproved of or disliked (Raihani & Bshary 2015b), people also disapprove of individuals who do not punish their peers (Martin et al., 2019). Even children sometimes prefer punishers to those who do not respond to wrongdoing (Loke et al., 2011, 2014; Vaish et al. 2016). A proposed key to positive evaluations is that punishment must be perceived as legitimate (Raihani & Bshary, 2019). As the everyday punishing events depicted in this study's vignettes should generally be perceived as a rightful action (e.g., demonstrating pro-social behaviour, justified self-protection, or enforcement of social norms), I expected the evaluation of punishers to be positive overall.

However, since punishment tends to be regarded more favourably when selfish or competitive motives can be discounted (Barclay, 2006; Jordan et al., 2016; Kiyonari & Barclay, 2008; Raihani & Bshary, 2015b, 2015c), and both second-party punishment and antisocial punishment are associated with competitive rather than cooperative motivations – which are more likely in TPP (Carpenter, 2003; Espin et al., 2012; Peysakhov et al., 2014; Yamagishi et al., 2012) -, my second hypothesis posited that third-party punishers would be evaluated more positively than second-party punishers.

Finally, based on the findings outlined above, I predicted that motive attributions would differ between punishing types. I for instance expected observers to associate prosocial motive items with TPP rather than SPP. Specifically, I anticipated that third-party punishers would be seen as motivated by a desire to protect or help others, and as willing to uphold the welfare of others, emphasizing principles of group living and fairness. In contrast, I expected self-protection, formidability and retribution motive items to be more strongly associated with SPP, as it involves more competitive and self-interested responses to wrongdoing. This includes motivations of appearing dominant, strong and authoritative, compared to TPP motives of appearing cooperative. I also expected observers to view motives of deterrence, defence of values, and the expression of disapproval as relevant to both second- and third-party punishers.

By investigating these hypotheses (see preregistration in Appendix D1), this study aimed to contribute to the understanding of the complex interplay between punishment, motives, and social evaluations.

## 5.1.2 Methods

### 5.1.2.1 Participants

87 participants completed the study, which was posted and re-posted on the subreddit r/SampleSize, as well as its German (r/samplesize\_DACH) and UK (r/SampleSizeUK) counterparts for a duration of two weeks (for an evaluation of Reddit as a crowdsourcing platform for psychology research projects see Luong & Lomanowska, 2022). After screening for suspicious looking responses (e.g. the same responses across all Likert-scale questions,  $n = 0$ ) and failed attention checks ( $n = 2$ ), four further participants were excluded for being under the age of 18, leaving 81 viable participants (29 male, 52 female). They were between 18 and 70 years old ( $M = 32.78$ ,  $SD = 13.22$ ) and were not compensated for their participation. Informed consent was obtained from all participants prior to data collection and the research was approved by the UCL Research Ethics Committee (Project ID: *ICN-PWB-NR-23-10-20A*). Study materials can be found in Appendix D2.

### 5.1.2.2 Procedure and Materials

First, participants provided their demographic information and then proceeded to respond to seven vignettes. After completing an attention check, they read the debriefing information, which included an optional comment box. Median completion time was nine-and-a-half minutes.

Each of the seven vignettes had two versions: one in which the punisher was the victim of a wrong-doing (SPP) and one in which someone other than the punisher had been the victim of a wrong-doing (TPP), while keeping all other elements consistent. For instance, the counterpart to the SPP vignette “Alex witnesses someone damaging *his* property and makes them compensate *him* for the damages” was “Alex witnesses someone damaging *his parent’s* property and makes them compensate *his parents* for the damages”. Participants were randomly presented with either the SPP or TPP version of each vignette, which were evenly presented (i.e., for each new vignette participants had an equal chance of seeing either the SPP or TPP version). The seven vignettes encompassed a diverse range of violations, as well as various relationships between the punishers, violators, and victims. Additionally, these vignettes incorporated a wide array of punishing methods, including physical confrontation, verbal confrontation, material punishment, gossip, exclusion, the reporting of violators, and

withdrawal of benefits. By encompassing this comprehensive range of scenarios and punishment strategies, I aimed to somewhat include the complexity and nuances of real-world punitive interactions. Specifically, scenarios, which were presented in random order, included reporting a disruptive gym member, no longer giving rides to a friend who stole ten pounds, hitting a mugger with a handbag, excluding a violator who made offensive comments from community events, gossiping about a co-worker who took credit for other people's achievements, and telling off a friend who took the last slice of someone else's pizza without asking (Table 5.1).

**Table 5.1**

*Second- and third-party punishment scenarios used in the study*

|   | SPP                                                                                                                                                                                                                                       | TPP                                                                                                                                                                                                                                                  |
|---|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | At the gym, Sarah's workout is consistently disrupted by Mark, another gym member. Sarah therefore reports Mark's behaviour to the gym, which revokes Mark's membership.                                                                  | At the gym, Sarah witnesses Mark consistently disrupting another gym member's workout. Sarah therefore reports Mark's behaviour to the gym, which revokes Mark's membership.                                                                         |
| 2 | As Jessica walks along the street she is mugged. However, Jessica yanks the handbag out of the would-be mugger's hands and, as he starts running away, hits him with it.                                                                  | As Jessica walks along the street she witnesses a mugging. However, Jessica rushes over, yanks the handbag out of the would-be mugger's hands and, as he starts running away, hits him with it.                                                      |
| 3 | At a community gathering a person repeatedly makes derogatory and offensive comments towards Dave. In response, Dave decides to exclude the person from future community events and gatherings.                                           | At a community gathering, Dave witnesses a person repeatedly making derogatory and offensive comments towards certain individuals. In response, Dave decides to exclude the person from future community events and gatherings.                      |
| 4 | Alex witnesses someone damaging his property and makes them compensate him for the damages.                                                                                                                                               | Alex witnesses someone damaging his parent's property and makes them compensate his parents for the damages.                                                                                                                                         |
| 5 | Emma becomes aware of a co-worker consistently taking credit for her achievements. So, Emma discreetly shares this information with other colleagues, leading to a decline in the other person's professional reputation and credibility. | Emma becomes aware of a co-worker consistently taking credit for other people's achievements. So, Emma discreetly shares this information with other colleagues, leading to a decline in the other person's professional reputation and credibility. |
| 6 | John discovers that Kim stole ten pounds from him and decides to stop giving her rides, which he usually does.                                                                                                                            | John discovers that Kim stole ten pounds from their friend and decides to stop giving her rides, which he usually does.                                                                                                                              |
| 7 | Mary and Jim are sharing an XL pepperoni pizza while Tamara has her own small vegetarian pizza. When Jim eats Tamara's last                                                                                                               | Mary and Jim are sharing an XL pepperoni pizza while Tamara has her own small vegetarian pizza. When Jim eats Tamara's last slice without asking,                                                                                                    |

slice without asking, she tells him off for being  
inconsiderate and selfish.

Mary tells him off for being inconsiderate and  
selfish.

---

After reading each vignette, participants responded to 18 motive items assessing the perceived motivations underlying the depicted punishing actions. These items encompassed a diverse range of motivations, spanning from the desire to gain reputation benefits by appearing prosocial (e.g., 'to appear cooperative,'), being genuinely prosocial (e.g., 'to be helpful,' 'to protect someone else'), appearing formidable (e.g., 'to show strength,' 'to gain respect or admiration') to motives of deterrence (e.g., 'to deter future misconduct,' 'to protect themselves') and retribution (e.g., 'to seek revenge,' 'to teach a lesson'). Motives related to public goods (e.g., 'to enforce rules or regulations,' 'to maintain social order') and conveying disapproval (i.e., 'to express disapproval') were incorporated as well. These specific motive items were included to capture the diverse range of motivations that participants may attribute to TPP and SPP actions, allowing for a comprehensive exploration of the underlying factors guiding punitive behaviours in the context of each vignette. Participants rated every motive item on a scale from 1 = *'Not at all'* to 7 = *'Completely'*, indicating the extent to which they believed the punisher in each vignette was motivated by the specified motive.

Participants also indicated whether they believed the punitive actions reflected more negatively or positively on the punisher by providing ratings on a scale from 1 = *'Extremely negatively'* to 7 = *'Extremely positively'*."

Lastly, to ensure participants' engagement, they were asked to complete an attention check by identifying the topic of the vignettes. The correct answer was *'responses to wrong-doings,'* while *'consumer purchasing habits,'* *'popular sports,'* and *'book preferences'* were provided as incorrect options.

### 5.1.2.3 Analysis Plan

All statistical analyses in this study were conducted in R (R Core Team, 2023) using the packages 'lme4' (Bates et al., 2015), 'lmerTest' (Kuznetsova et al., 2017), and 'emmeans' (Lenth, 2023). All reported coefficients are unstandardized. All models were estimated using REML and nloptwrap optimizer, and mixed model Analyses of Variance (ANOVAs) with Satterthwaite's method were conducted for the second and third models described below.

To examine participants' evaluations of punishers and identify below or above average effects, the punisher evaluation scores were centred around the scale midpoint of 4. This adjusted evaluation score was then employed as the response variable in a constant (intercept-only) linear mixed model, incorporating participant ID as a random effect.

To investigate whether third-party punishers receive more positive evaluations compared to second-party punishers, a linear mixed effects model was fit. Participants' evaluations of punishers were used as the response variable, while punishing type (1 = TPP, 0 = SPP) was included as a fixed effect, and participant ID was specified as the random effect.

Although it was inadvertently overlooked during pre-registration, I decided to include Story ID as a control prior to data analysis, because several context factors may influence motivation ascriptions (see Appendix D3 for the analysis without Story ID as control). Details of social relationships may affect responses across vignettes, as observers for instance believe it is more appropriate for those who are socially closer to the violator to confront them (Eriksson et al., 2017; Strimling & Eriksson, 2014). Moreover, observers take the relative cost of punishment into account when evaluating punishers (Nelissen, 2008). As direct punishment is often more costly than indirect punishment (Archer & Coyne, 2005; Molho et al., 2022; Nikiforakis, 2008), scenarios involving direct punishing methods (e.g., hitting a mugger) might be evaluated more favourably than more indirect actions (e.g. refraining from giving rides) or scenarios in which the punisher may otherwise benefit from engaging in punitive actions (e.g. even when demanding compensation as a third-party punisher 'Alex' could receive financial benefits downstream). On the other hand, direct confrontations can have the opposite effect if they are perceived as aggressive and linked to selfish motives (Eriksson et al., 2016; Kupfer & Giner-Sorolla, 2017). As it is possible for interaction effects to exist (e.g., it may be more appropriate to defend oneself against a mugger than attack as a third-party, whilst gossiping may be viewed as more prosocially motivated when carried out by an unaffected co-worker versus spitefully when carried out by an affected individual), an interaction between punishing type and Story ID was included.

In order to explore the associations between motive items and punishing types, another linear mixed effects model was applied. The motive score was treated as the response variable, and both motive item and punishing type were included as fixed effects, with an interaction term between them. Participant ID was again included as a random effect to account for individual variability, and Story ID was again retrospectively added as a control.



### 5.1.3 Results

Comparing participants' evaluations of punishers to the scale midpoint revealed that punishers were generally viewed favourably,  $b = 1.02$ ,  $t(80) = 9.75$ , 95% CI [0.81, 1.22],  $p < .001$ .

The model examining how punishers are judged depending on punishing type (SPP versus TPP), punishing story (i.e., with varying context factors such as punishing method, violation, relationships) and their interaction had a substantial total explanatory power (conditional  $R^2 = 0.49$ ). Both story ( $F(6, 10108) = 326.77$ ,  $p < .0001$ ) and punishing type ( $F(1, 10142) = 81.55$ ,  $p < .0001$ ) significantly influenced evaluations of punishers, as did their interaction ( $F(6, 10139) = 14.52$ ,  $p < .0001$ ). Punishers were generally judged significantly more favourably when engaging in TPP compared to SPP.

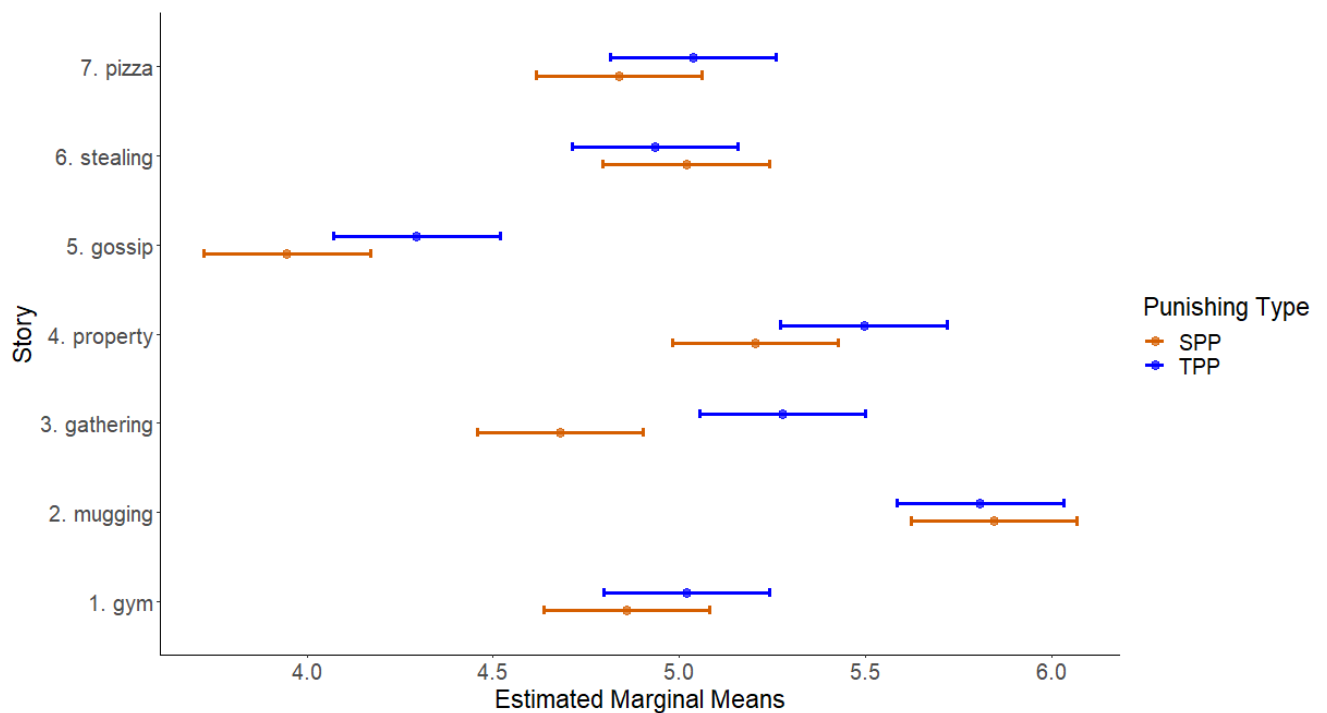
Overall, participants indicated that the scenarios in which punitive actions most positively reflected on the punisher were: hitting a mugger with a handbag ( $M = 5.83$ , 95% CI [5.61, 6.04]) and demanding compensation for property damage ( $M = 5.35$ , 95% CI [5.14, 5.57]). This was followed by the vignettes in which the punisher engages in withdrawal of benefits (no more car rides) because the violator stole £10 ( $M = 4.98$ , 95% CI [4.76, 5.19]), excluding the violator from community gatherings for making derogatory comments ( $M = 4.98$ , 95% CI [4.76, 5.19]), reprimanding a friend for eating the last slice of someone else's pizza without asking ( $M = 4.94$ , 95% CI [4.72, 5.15]), and reporting a disruptive gym member ( $M = 4.94$ , 95% CI [4.72, 5.15]). Engaging in workplace gossip as a way to punish someone who takes credit for others' work was regarded least positively ( $M = 4.12$ , 95% CI [3.91, 4.34]).

Due to the significant interaction effect, simple pairwise comparisons using the emmeans package were conducted. First, estimated marginal means were calculated for the interaction between punishing type and story, using the emmeans function, followed by pairwise comparisons using the pairs function. A Bonferroni correction was applied to adjust for multiple comparisons, with the new significance level set at  $\alpha < .007$  (i.e.,  $0.05/7$ ). Figure 5.1 illustrates that TPP was consistently evaluated as reflecting significantly more positively on punishers than SPP across stories. However, there was no significant difference for the scenario in which a mugger is hit with a handbag ( $b = 0.04$ ,  $p = 0.55$ ), the scenario in which future favours of car rides are withheld to punish theft ( $b = 0.08$ ,  $p = .17$ ), or the scenario in which a disruptive gym member is reported ( $b = -0.16$ ,  $t = -2.61$ ,  $p = .009$ ). The more favourable view of third-party punishers was most pronounced in the scenario in which someone making derogatory remarks is excluded from community gatherings ( $b = -0.60$ ,  $t = -9.73$ ,  $p < .0001$ ), followed by the

scenario in which a co-worker taking credit for others work is gossiped about ( $b = -0.35$ ,  $t = -5.68$ ,  $p < .0001$ ), the scenario in which compensation for property damage is demanded ( $b = -0.29$ ,  $t = -4.74$ ,  $p < .0001$ ), and lastly the scenario in which a friend takes the last slice of someone else's pizza without asking ( $b = -0.20$ ,  $t = -3.23$ ,  $p = .001$ ).

**Figure 5.1**

*Estimated marginal means showcasing the interaction between Story and Punishing Type*



*Note.* Dots represent estimated marginal means and error bars signify 95% confidence intervals (CIs).

The model examining which motive items were differentially relevant for SPP compared to TPP, controlling for story, revealed significant main effects for punishing type ( $F(1, 10158) = 43.924$ ,  $p < .0001$ ), motive item ( $F(17, 10080) = 209.821$ ,  $p < .0001$ ), and story ( $F(6, 10080) = 30.907$ ,  $p < .0001$ ), as well as a significant interaction effect between punishing type and motive item ( $F(17, 10080) = 91.751$ ,  $p < .0001$ ).

Post-hoc analyses were conducted to examine the interaction between punishing type and motive item on motive scores. First, estimated marginal means were calculated for the interaction between punishing type and motive item while controlling for story. Subsequently,

pairwise comparisons of the estimated marginal means for punishing type and motive item were conducted. A Bonferroni correction was applied to adjust for multiple comparisons, with the new significance level set at  $\alpha < .0028$  (i.e.,  $0.05/18$ ).

Several motives were associated significantly more with TPP than SPP. Most strongly, participants attributed motives to protect someone else ( $b = -3.14, t = -23.03, p < .0001$ ) and to be helpful ( $b = -2.58, t = -18.95, p < .0001$ ) to third-party punishers over second-party punishers. Participants also thought third-party punishers were more motivated by wanting to appear cooperative ( $b = -1.11, t = -8.13, p < .0001$ ), maintaining social order ( $b = -0.82, t = -6.01, p < .0001$ ), and defending personal or collective values ( $b = -0.47, t = -3.43, p = .0006$ ).

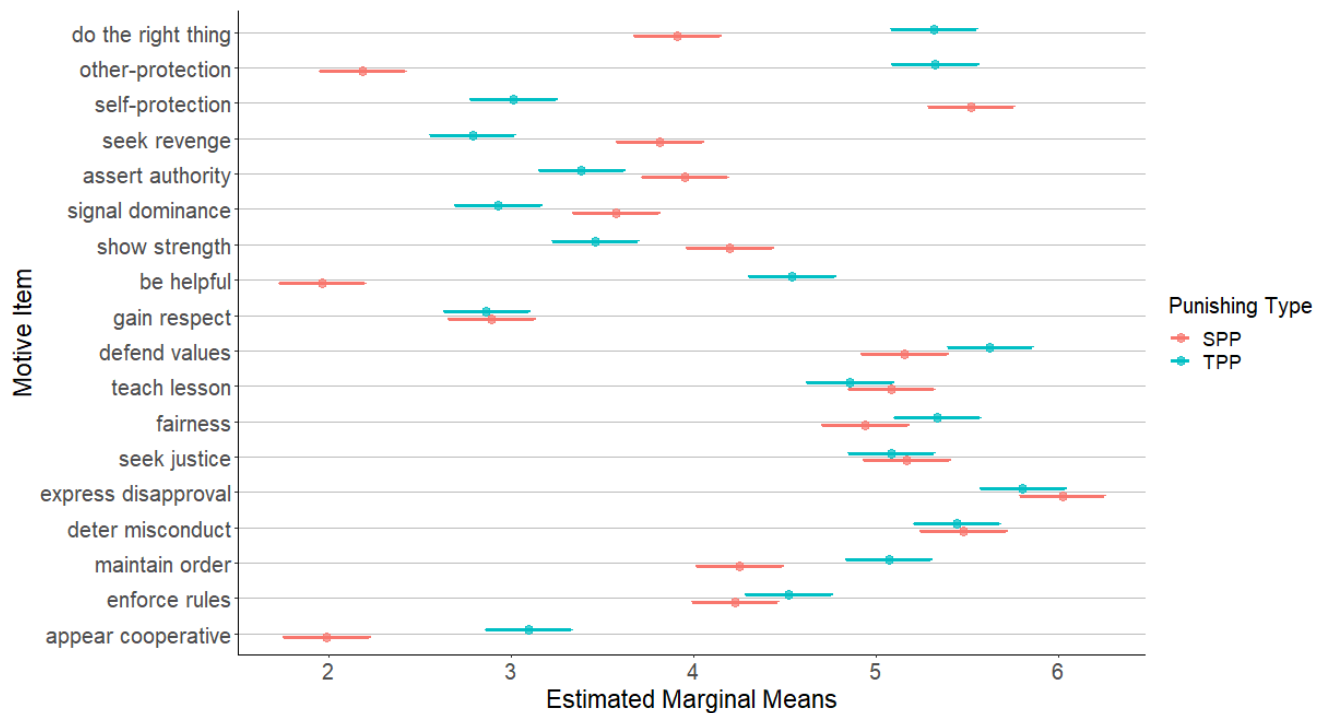
The motives that were most strongly associated with SPP compared to TPP were self-protection ( $b = 2.51, t = 18.44, p < .0001$ ), revenge ( $b = 1.03, t = 7.56, p < .0001$ ), showing strength ( $b = 0.74, t = 5.42, p < .0001$ ), signalling dominance ( $b = 0.65, t = 4.75, p < .0001$ ) and asserting authority ( $b = 0.57, t = 4.17, p < .0001$ ).

Motives for which no significant difference was found between SPP and TPP included gaining respect or admiration ( $b = 0.03, t = 0.22, p = .83$ ), deterring future misconduct ( $b = 0.035, t = 0.26, p = .80$ ), seeking justice ( $b = 0.083, t = 0.61, p = .54$ ), expressing disapproval ( $b = 0.22, t = 1.59, p = .11$ ), teaching a lesson ( $b = 0.23, t = 1.67, p = .09$ ), enforcing rules or regulations ( $b = -0.29, t = -2.15, p = .03$ ), and restoring balance or fairness ( $b = -0.40, t = -2.91, p = .004$ ).

As can be seen in Figure 5.2, participants believed punishment mostly served as a means to express disapproval ( $M = 5.91, 95\% \text{ CI } [5.72, 6.10]$ ) and to deter future misconduct ( $M = 5.46, 95\% \text{ CI } [5.27, 5.56]$ ). In contrast, observers did not think that second- and third- punishers were strongly motivated by a desire to appear cooperative ( $M = 2.54, 95\% \text{ CI } [2.35, 2.73]$ ) or gain respect and admiration ( $M = 2.88, 95\% \text{ CI } [2.69, 3.07]$ ).

**Figure 5.2**

*Estimated marginal means demonstrating relevance of motive items across punishing type*



*Note.* Dots represent estimated marginal means and error bars signify 95% confidence intervals (CIs).

Lastly, the relationships between motives were examined using correlation analyses. For second-party punishment scenarios, the weakest correlations were between the motives deterring future misconduct and appearing cooperative ( $r = 0.00$ ), expressing disapproval and seeking revenge ( $r = -0.01$ ), protecting oneself and being helpful ( $r = -0.01$ ), deterring misconduct and signalling dominance ( $r = 0.01$ ), gaining respect and doing the right thing ( $r = 0.02$ ), seeking justice and protecting another ( $r = 0.02$ ), seeking justice and being helpful ( $r = 0.02$ ), deterring future misconduct and being helpful ( $r = -0.02$ ), maintaining order and seeking revenge ( $r = -0.02$ ), as well as gaining respect and protecting oneself ( $r = -0.02$ ). Conversely, the strongest correlations between SPP motives were asserting authority and signalling dominance ( $r = 0.65$ ), signalling dominance and showing strength ( $r = 0.59$ ), protecting another and being helpful ( $r = 0.58$ ), asserting authority and showing strength ( $r = 0.52$ ), as well as being helpful and appearing cooperative ( $r = 0.52$ ). See Appendix D4 for the full correlation matrix.

For third-party punishment scenarios, the weakest correlations were between the motives showing strength and deterring future misconduct ( $r = 0.00$ ), showing strength and doing the right thing ( $r = 0.00$ ), restoring fairness and seeking revenge ( $r = 0.00$ ), teaching a lesson and protecting oneself ( $r = 0.00$ ), gaining respect and defending values ( $r = 0.01$ ), expressing disapproval and protecting oneself ( $r = 0.01$ ), restoring fairness and signalling dominance ( $r = 0.01$ ), protecting another and showing strength ( $r = 0.01$ ), as well as teaching a lesson and being helpful ( $r = 0.01$ ). Conversely, the strongest correlations between TPP motives were asserting authority and signalling dominance ( $r = 0.72$ ), asserting authority and showing strength ( $r = 0.68$ ), signalling dominance and showing strength ( $r = 0.66$ ), gaining respect and asserting authority ( $r = 0.63$ ), gaining respect and signalling dominance ( $r = 0.62$ ), protecting another and being helpful ( $r = 0.58$ ), gaining respect and showing strength ( $r = 0.55$ ), doing the right thing and protecting another ( $r = 0.54$ ), as well as restoring fairness and seeking justice ( $r = 0.53$ ). See Appendix D4 for the full correlation matrix.

### 5.1.4 Discussion

This vignette study aimed to fill a gap in the existing literature by examining the evaluation of and motives attributed to adult peer punishers in both third-party and second-party contexts. Each vignette illustrated a distinct scenario involving a wrongdoing and subsequent punishment. The vignettes encompassed a range of direct and indirect punishing responses to different norm violations, improving the ecological validity of the current findings (Molho et al., 2020; Molho & Wu, 2021).

All three hypotheses were largely supported. Firstly, I found that punishers were generally viewed favourably. Consistent with the idea that perceptions of legitimacy play a crucial role in shaping positive evaluations of punishers (Raihani & Bshary, 2015b), participants likely perceived the punitive actions depicted in the vignettes as legitimate. For instance, they may have attributed prosocial or justified self-protection intentions to the punishers.

According to previous research (Barclay, 2006; Jordan et al., 2016; Kiyonari & Barclay, 2008; Raihani & Bshary, 2015b,c), punishment tends to be viewed more favourably when selfish or competitive motives can be discounted. This is less likely during SPP than during TPP events, as punishers are observers rather than directly affected during TPP. In line with this, previous research has found that SPP is associated with competitive rather than cooperative motivations, whereas TPP is more likely to be driven by cooperative intentions (Carpenter, 2003; Espin et al., 2012; Jordan et al., 2016; Peysakhov et al., 2014; Yamagishi, 2012). Matching these prior studies and my predictions, the present study also found that TPP was regarded more favourably than SPP.

Specifically, the preference for TPP over SPP was most pronounced for the vignette in which the punisher excluded someone from future gatherings in response to someone else being insulted, rather than the punisher themselves being insulted. Instances in which SPP was viewed as equally acceptable to TPP were the scenario in which the punisher hit a mugger with a handbag, and the scenario in which the punisher stopped giving his friend rides after she stole money. A reason for this may be that individuals would not typically expect someone to do favours for someone who has wronged them. Furthermore, physically confronting a mugger might be seen as a situation where SPP is particularly acceptable and not perceived as driven by (antisocial) selfish or competitive motives. In both cases, the punishments are likely considered rightful, whether motivated by concern for oneself or for others, which, as mentioned above, is essential for positive evaluations of punishers (Raihani & Bshary, 2019).

Overall, the vignettes in which punishment received the most favourable evaluations were when punishers engaged in physical confrontation (the mugging scenario) and demanded material compensation for property damage. Again, punishment was perhaps viewed as most legitimate or least motivated by selfish or competitive concerns in these two scenarios. Additionally, the mugging vignette arguably presented the most costly punitive act due to the inherent risk of physically attacking a mugger. Punishers who are willing to pay high costs are regarded more positively because selfish motives for punishing appear less likely (Nelissen, 2018). Gossip, which as an indirect punishing method is less costly, was regarded least favourably. This punishing method may have been regarded as especially unsuitable because the punished may not learn about or from the punishment, excluding motives of specific deterrence. However, these are merely conjectures. For a reliable account see e.g. Molho et al. (2020), who investigated direct and indirect punishment in daily life. These interaction effects for story further showcase the importance of context in evaluating punishers.

Previous research has shown that punishers can be motivated by various factors, including revenge (Bone & Raihani, 2015; Deutchman et al., 2021; Twardawski et al., 2022), status concerns (Redhead et al., 2021; Yamagishi et al., 2012), aversion to (disadvantageous) inequality (Bone et al., 2016; Bone & Raihani, 2015; Raihani & McAuliffe, 2012), a desire to communicate disapproval (Aharoni et al., 2022; Cushman et al., 2019; Funk et al., 2014; Sarin et al., 2021), deterrence of future harmful behaviour (Delton & Krasnow, 2017; McCullough et al., 2013), as well as competitive motives (Bone & Raihani, 2015; Price et al., 2002; Raihani & Bshary, 2019). However studies that directly investigate the differential attribution of motives by observers to second- and third-party punishers is scarce.

Since TPP is administered by uninvolved bystanders and has frequently been associated with cooperative intentions (e.g. Jordan et al., 2016; Jordan & Rand, 2017; Raihani & Bshary, 2015b,c), I anticipated that observers would attribute prosocial motive items to third-party punishers. These motives encompass desires to cultivate a prosocial reputation, protect others, and promote group cohesion by upholding social norms, which in turn enhances group-level cooperation (Boyd & Richerson, 1992; Gintis et al., 2008). In contrast, SPP has been linked to competitive and self-centred motives (Carpenter, 2003; Peysakhov et al., 2014; Raihani & Bshary, 2015b; Yamagishi et al. 2012) leading me to expect observers to associate formidability and retribution items to second-party punishers. Observers were expected to perceive motives such as deterrence, defending values, and communication of disapproval as relevant for both SPP and TPP.

In line with predictions, the current study found that observers did identify certain motive items as differentially relevant for SPP and TPP. Participants believed that third-party punishers were distinctly motivated by a desire to protect someone else, to be helpful, appear cooperative, maintain social order, and defend personal or collective values. On the other hand, they attributed motives of self-protection, revenge, showing strength, signalling dominance and asserting authority to second-party punishers over third-party punishers. They appeared to find motivations of gaining respect or admiration, deterring future misconduct, seeking justice, expressing disapproval, teaching a lesson, restoring balance, and enforcing rules as equally likely for second- and third-party punishers. Importantly, observers believed that punishers were most driven by a desire to express disapproval and to deter future misconduct, whereas they rejected the notion that punishers wanted to appear cooperative or gain respect and admiration.

Contrary to predictions, attributions of defending collective or personal values were more commonly associated with third-party punishers rather than second-party punishers. One possible reason for this could be that participants did not perceive second-parties as frequently or strongly motivated by the defence of either personal or collective values when engaging in punishment. In contrast, they may have believed that third-parties aimed to uphold social norms by defending collective values. Furthermore, although third-party punishers were significantly more likely to be attributed with a desire to appear cooperative compared to second-party punishers, overall, this motive received the least endorsement. The motive of gaining respect was also minimally endorsed. These findings suggest that, at least consciously, observers do not appear to expect punitive actions to be primarily motivated by a desire to gain reputational benefits.

However, I did not investigate the diverse reputational consequences of punishment beyond evaluating whether participants perceived the punitive action as reflecting positively or negatively on the punisher, nor did I include explicitly competitive motives. Since the potential for punishers to enhance their fitness or status relative to others is an important function of punishment (Bone & Raihani, 2015; Deutchman et al., 2021; Price et al., 2002; Raihani & Bshary, 2019), it would be beneficial for future research to explicitly include this aspect for a more comprehensive understanding on the perception of punishers motives. It would additionally be compelling to assess the extent to which each punitive action appeared to be motivated by self-centred or other-regarding concerns (Raihani & Bshary, 2015b).



Furthermore, it would be interesting to directly investigate whether observers motive attributions and punishers actual reported motives align.

In summary, the present study found that punishers were generally viewed favourably, and that third-party punishers were evaluated more positively than second-party punishers. Most importantly, the study demonstrated that punishing motives are differentially perceived depending on the type of punishment. Observers associated TPP with prosocial motivations, including the promotion of cooperation, whereas SPP was associated with self-protection and formidability concerns. Overall, participants considered deterrence and communications of disapproval to be the most important drivers of punishment (highly relevant to both SPP and TPP), whilst they did not consider reputation management strategies to be particularly relevant. Collectively, these findings offer valuable insights into the intricate ways in which individuals perceive and interpret the act of punishing transgressors, a fundamental aspect of human behaviour.

## **5.2 Meta-Cognitive and Observer Accuracy Surrounding Punisher Motive Attributions**

### **5.2.1 Introduction**

In the first study of this Chapter I used vignettes to investigate the motives observers attribute to second- and third-party punishers, as well as observers opinions on whether punitive responses reflected positively or negatively on punishers. While some motives were equally attributed to second- and third-party punishers (e.g., the two most attributed motives ‘expressing disapproval’ and ‘detering future misconduct’), I found that observers associated TPP more with prosocial motivations and SPP more with self-protection and formidability concerns. Third-party punishers also received more reputation benefits than second-party punishers, with observers believing that punitive TPP actions reflected more positively on punishers than equivalent SPP actions.

However, it would be interesting to see whether these findings can be replicated when a more complex method is used. Although the vignettes covered a range of direct and indirect punishing responses to violations applicable to everyday life, eliciting actual real life instances of punishment would provide even more ecological validity (Molho et al., 2020). Instead of ensuring experimental control by manipulating only the punishing context (SPP versus TPP) of each vignette, this study directly sampled punitive responses to wrong-doings from peoples’ everyday lives. This method ensured that a larger range of punishing methods and contexts (e.g. with various social relationships or severity levels of violation and punishment) were provided, which is more reflective of our complex real-world environments.

In addition to investigating whether this method results in similarly distinct punishing motive attributions as observed in the vignette study, the present study investigated whether punishers themselves believe their actions result from the same distinct motives. More importantly, as individuals not only consider the potential direct consequences of punishing another, but also how their punitive action will influence observers’ judgments about them and their underlying motives when deciding whether to punish (Radkani et al., 2022), a main objective for this study was to investigate whether people know which motives observers are likely to ascribe to their punishing actions (meta-accuracy). Conversely, another main objective was to examine

whether observers' motive attributions match punishers' actual reported motives (observer accuracy). To determine observer- and meta-accuracy I collected self-perceptions and meta-perceptions from punishers, as well as actual perceptions from observers (see Appendix D5 for a preregistration of this study and Appendix D6 for its study materials).

People expect punishment to be both produced and understood as a communicative action (Cushman et al., 2019; Funk et al., 2014; Ho et al., 2019, 2021; Molnar et al., 2020; Sarin et al., 2021). Conveying and inferring communicative intent relies on the ability to model other individual's minds (Baker et al., 2009, 2017; Wellman, 1992; Young & Saxe, 2009). Both the actor and observer must therefore model the relationship between the actor's actions and their possible motives, as well as the interpretations observers may make (Sarin et al., 2021). For punishment to be a successful communicative action, actor meta-perceptions, observer motive attributions and actors' actual motives have to align.

Punishment is sometimes favoured because it can improve a punisher's status or situation (Deutchman et al., 2021; Raihani & Bshary, 2019), but it can also provide opportunities for receiving reputation benefits – especially in the case of TPP (Mifune et al., 2020). Although observers in the vignette study rejected the notion that punishers wanted to 'appear cooperative' or 'gain respect and admiration', this study explored whether actors nevertheless gain reputation benefits for their punishing actions, and whether they gain distinct reputation benefits depending on whether punishment occurred in a second- or third-party punishing context. As punishers tend to be regarded more favourably when selfish or competitive motives can be discounted (Barclay, 2006; Jordan et al., 2016; Kiyonari & Barclay, 2008; Raihani & Bshary, 2015b,c), and cooperative motivations are more likely when the punisher was not the victim of the violation (Carpenter, 2003; Espin et al., 2012; Peysakhov et al., 2014; Yamagishi et al., 2012), I expected third-party punishers to receive more favourable assessments in the realm of cooperative and prosocial reputation than second-party punishers. Although punishers in both contexts can desire and achieve a reputation for being formidable, this is likely the most relevant reputation domain for second-party punishers. Therefore, I expected punishers and observers to attribute more formidable traits to second- compared to third-party punishing situations. Overall, I expected punishers to be evaluated more positively in TPP compared to SPP contexts, and for TPP to be perceived as more justified. However, as punishment signalling is sensitive to deservedness (Salcedo & Jimenez-Leal, 2023), as also demonstrated by my findings in Chapter 3, I expected positive evaluations of punishers to be more strongly influenced by the perceived justness of a punishing action than by the type of punishment.

Similar to Chapter 4, to assess observer and actor meta-perceptive accuracy I again used a componential approach adapted from the Social Accuracy Model (SAM; Biesanz, 2010; Lees et al., 2022). With this adaptation I could determine baseline accuracy, normative accuracy, distinctive accuracy, transparency bias (e.g., whether punishers overestimate how transparent their attributes are to others), meta-insight (e.g., whether punishers know how they will be misperceived), systematic meta-bias, as well as trait moderators of actor and observer accuracy. In contrast to Chapter 4, which mainly investigated whether actors know how their cooperative and uncooperative behaviours affect their reputation, this study focused equally on observer accuracy and actor meta-accuracy.

Observers use the punishing context to infer what motivated a punishment decision, from which they extrapolate how punishers are likely to behave in future (Raihani & Bshary, 2015b). However, because punishment often improves the punisher's situation compared to the wrongdoers', it has been suggested that punishers are generally met with observers' doubt regarding the moral legitimacy of the punishment (Fehr & Rockenback, 2003; Raihani & Bshary, 2015b; Xiao, 2013). Even when punishers were not the victim of the wrong-doing, observers may believe that the punisher acted out of jealousy or for competitive reasons (Dhaliwal et al., 2020; Pedersen et al., 2013). Due to the complexity of a punisher's reputation, it may therefore be difficult for observers and punishers to provide accurate ratings. On the other hand, research suggests that punishment in the real world is not a common occurrence (Balafoutas et al. 2014; Baumard 2010; Guala 2012; Pedersen et al. 2018): perhaps people are only likely to punish when they are fairly certain about how their action will be perceived. As people are generally accurate at estimating how others perceive them across social contexts (Carlson et al., 2010, 2011; Carlson & Furr, 2009; Carlson & Kenny, 2012; Carlson et al., 2011; Kenny & DePaulo, 1993; Shechtman & Kenny, 1994), and are able to accurately anticipate the reputation consequences of compensation versus punishment (Dhaliwal et al., 2020), I again expected participants to provide accurate observer and meta-perceptions for both second- and third-party punishing behaviours.

Furthermore, I anticipated that this study could reveal punishing type as a moderator of meta-perceptive and observer accuracy. Accuracy could be enhanced for TPP contexts, because third-party punishers are more likely than second-party punishers to be regarded positively for their punitive actions (Mifune et al., 2020; Raihani & Bshary, 2015b), making it more important to accurately predict how punishing actions will be perceived in TPP contexts. Even third-party punishers can gain a reputation for being formidable resulting in lower chances of future harm

(Krasnow et al., 2012, 2016), but to my knowledge there are currently no studies which show that second-party punishers benefit from receiving a prosocial reputation. On the other hand, accuracy could be enhanced for SPP contexts. SPP is more common than TPP as second parties are more likely to seek out opportunities to punish whereas third parties punish reluctantly (Kriss et al., 2016). With more frequent exposure to SPP, people may be more practiced in accurately gauging how their actions are perceived by others in SPP contexts. Moreover, motives for SPP could be regarded as more straight forward (e.g. revenge, competition) whereas seemingly prosocial TPP actions could have additional underlying motives. Altogether, this invited an exploration of the possible differences in observer and meta-perceptive accuracy between TPP and SPP.

Whenever punishers misestimate how they are perceived, I expected them to believe others would judge them less favourably than they actually do, indicating a systematic negative meta-bias. Knowing whether a punishing act will be perceived as morally legitimate is important for the punisher, because if observers conclude it was not legitimate, this increases retaliation risks and is less likely to deter future wrong-doing (Raihani & Bshary, 2019; Xiao, 2013). The cost of a positivity bias (believing their punishing action will be regarded more favourably than it actually is) is therefore high. A negativity bias would be comparatively less costly, as individuals may miss out on opportunities for gain, but they also do not risk negative consequences (for an exception of higher-order punishment see Martin et al., 2019). Additionally, there is evidence to suggest that punishing decisions are guided by the perceived norm of punishment (Dhaliwal et al., 2020; Fabbri & Carbonara, 2017; Yang & Chen, 2022), specifically that people underestimate the norm of punishment which lowers levels of punishment (Dhaliwal et al., 2020; Yang & Chen, 2022). As individuals underestimate how normative punishment is, perhaps they therefore also believe it is less approved of than it is, making it more likely that participants will systematically underestimate how positively they are viewed for their punishing behaviours.

Lastly, I expected to find trait moderators for distinctive meta- and observer accuracy. As punishing only bestows reputation benefits in certain situations (e.g., Nelissen, 2008; Raihani & Bshary, 2015b), I anticipated that concern for reputation would positively moderate meta-accuracy. Individuals who are more concerned about their reputation may pay more attention to observers. Such an active awareness and motivation to understand perceivers could result in greater meta-accuracy. Furthermore, I expected those who have a higher propensity to engage in social perspectives taking or have a higher empathy quotient to exhibit enhanced meta-

perceptive or observer accuracy. These skills may be necessary to accurately predict how others will view them, or to accurately assess what really motivates another's behaviour. Indeed, as already outlined in Chapter 4, Tsankova & Tair (2021) suggested that a positive association between social competencies and meta-accuracy is possible, as there are reports of impaired meta-accuracy in autism (Sasson et al., 2018). I also expected participants who are highest on the subjective social status ladder to demonstrate lower accuracy than participants ranking lowest on the social status ladder. Accurate observer and meta-perceptions may be especially useful for individuals of lower-status, as it is particularly important for lower-status individuals to anticipate outcomes which higher-status individuals control (Fiske, 1993; Snodgrass et al., 1998). Similarly, individuals of higher status tend to pay less attention to lower-status individuals (Goodwin et al., 2000), which could increase the likelihood of those individuals missing cues necessary for accurate perceptions. Indeed, prior research has shown that individuals of lower status tend to have higher meta-accuracy than higher-status individuals (Carlson, 2016). However, another recent study found that people with low SES incorrectly believed that they are viewed more negatively than individuals with high SES (Engstrom et al., 2023). As recent research observed that individuals with strong meta-perception skills in first impressions tend to be extroverted (Hater et al., 2023), I also investigated whether personality traits, such as extroversion, moderate accuracy.

## **5.2.2 Methods**

### **5.2.2.1 Sampling Plan**

Based on the sampling plan outlined in Chapter 4, I again aimed to collect 140 usable TPP behaviour descriptions and 140 usable SPP behaviour descriptions by oversampling by 40 participants, resulting in a goal to recruit 180 actor participants. Again, based on the decision to have at least 10 judges rate each story, and an additional trade-off between a high number of observations and quality of responses (meaning each observer was asked to rate only 8 stories), 350 observers would be needed so that each of the anticipated 280 descriptions would be rated by around 10 observers. In anticipation of exclusions, I again aimed to oversample by 30 participants, resulting in 380 observer participants.

As I collected 294 usable stories from 147 actors, I recruited 398 observers to ensure that the additional behaviour descriptions were also rated by around 10 observers each. Further, 182 rather than 180 actor participants were recorded, as two actors were marked as “timed out” by Prolific despite them completing the study in a timely manner.

### **5.2.2.2 Participants**

A total of 580 participants were recruited from Prolific Academic in two waves. Only those who were US American, English-speakers and had an approval score of 100 were invited to take part in these studies. In addition, I used Prolifics’ “Quota Sampling” feature, which aims for an even gender balance. First, 182 “actors” (92 male, 90 female; after exclusions 75 male, 72 female) between the ages of 18-70 years ( $M = 37.26$ ,  $SD = 11.82$ ) were asked to describe a recent TPP and a recent SPP action they engaged in, before providing self- and meta-perceptions. Actor participants were predominantly White (65%). After careful consideration responses from 35 participants were excluded from further study for not describing a SPP/TPP action they engaged in ( $n = 20$ ; e.g., for describing TPP when SPP was asked for and vice versa), for being unspecific ( $n = 1$ ), or for including sensitive content such as murder, rape, child and animal abuse ( $n = 14$ ). None of the actors failed the attention check or more than two comprehension questions. Consequently, behaviour descriptions from 147 actor participants were presented to 398 observer participants, who each read and judged a random selection of 8 descriptions. 8 observers failed at least one of the attention checks and were therefore removed from all analyses. The remaining observers (195 male, 195 female) were between 21-75 years

old ( $M = 44.53$ ,  $SD = 12.74$ ) and predominantly White (71%). Actors were not able to participate as observers. The median completion time was 18 minutes for actors and 20 minutes for observers. They were paid at a rate of £9 per hour. All participants provided informed consent and the research was approved by the UCL Research Ethics Committee (Project ID: *ICN-PWB-NR-23-10-20A*).

### **5.2.2.3 Procedure**

#### ***Stage 1 of Data Collection - Actor Participants***

As in Chapter 4, participants from the first stage of data collection were explicitly asked for permission to use their descriptions as stimuli in future research after providing general informed consent. Similarly to the procedure outlined in Chapter 4, actors were told that this study aimed to investigate whether people can accurately predict how others will judge them based on their real-life behaviours. Actors were informed that they would take part in the first stage of this study, in which they would be asked to describe two recent punishing behaviours, and to predict how others will think of them based on the descriptions. Further, it was explained that in the second stage another set of participants would read their behaviour descriptions and rate how they view the participant based on the behaviour. Participants who would not want their descriptions to be used as stimuli in future research were asked to return the study on Prolific.

The remaining participants were asked to read the following information carefully, and were warned that their submission would be rejected if they repeatedly failed the comprehension checks. It was explained that they would be asked to describe two distinct punitive actions they had recently taken: one in which they punished another person when they themselves were the victim of an offense, and another in which they punished a wrong-doer after observing the offence as a third party. Similarly to the background explanation provided in Chapter 3, actors read the following information across several pages: *“In our daily lives we punish people to varying degrees. The wrong-doings that we punish can be serious (e.g. stealing) or mundane (e.g. making an inconsiderate comment), and can be committed by someone familiar or a stranger. Punishment is a behaviour that inflicts a cost on (or removes benefits from) a wrong-doer in response to that persons wrong behaviour. Punishment can take many forms, such as directly confronting a wrong-doer (either physically or verbally), sharing negative information about the wrong-doer to others (i.e., gossiping), demanding monetary compensation, discipline*



(e.g., grounding a child), excluding the wrong-doer from social interactions or avoiding social contact with them.” To ensure participants provided distinct second- and third-party punishment behaviour descriptions, it was emphasised that they could either punish a wrong-doer after they wronged them, or after they wronged someone else. An example of the former was to “*stop giving your friend rides because they stole \$50 from you*” and an example of the latter to “*stop giving your friend rides because they stole \$50 from another person*”.

To ensure that participants understood what punishment entails, particularly the distinction between second- and third-party punishment, actor participants were presented with four comprehension check questions. If participants answered a question incorrectly, they were told so, and given further explanations. If they answered correctly they saw the word “correct” in green before proceeding. First, participants were asked to pretend that they had written the punishing behaviour description: “*I caught my ex vandalising my car, so I demanded they pay me for the damages.*” Then, they were asked to indicate which response option matches the description, with “*punishment on behalf of myself*” being the correct response. The correct response to the second question “*Which of the behaviours below is a description of punishment on behalf of someone else*” was “*I saw a girl being harassed at the gym so I reported the harasser*”. The third question asked participants to identify which response matched a description of punishment on behalf of themselves. The correct answer was “*A coworker was taking credit for my work so I told other people about it to damage the coworkers reputation*”, whilst the incorrect options were “*I was bored so I went to annoy my sibling*” (because it does not involve a wrong-doing that is being punished) and “*Both describe punishment of a wrong-doing on behalf of myself*”. The last comprehension question asked “*Which of the following behaviours below is a description in which you punished someone on behalf of someone else*”. Hereby, the incorrect options described the participant as the wrong-doer, someone else punishing the wrong-doer on the participants behalf, and the participant punishing on behalf of themselves – making “*none of the above*” the correct response.

Subsequently, participants were asked to provide a written description of a recent TPP or SPP behaviour – the order in which they were asked to describe the punishing behaviours was again counterbalanced. The prompt for both behaviours was similar to that described in the previous Chapter, except that the part pertaining to cooperative or uncooperative behaviours now instructed participants to describe punitive actions (i.e., a recent instance in which they punished someone for engaging in a wrong behaviour that affected *them* and a recent instance in which they punished someone for engaging in a wrong behaviour that affected *someone else*).

Self- and meta-perceptions provided for participants' SPP and TPP actions were displayed as in Chapter 4, with the exception of which specific items were included. Although the preceding vignette study of the present Chapter included items linked to competitive motives (e.g. showing strength, signalling dominance, asserting authority, gaining respect/admiration), I wanted to include a further item directly assessing competitiveness as a motive. Moreover, I wanted to assess the extent to which each punitive action appeared to be motivated by self-centred or other-regarding concerns. This study therefore not only included all 18 previously used motive items, but introduced five additional ones to create an even more comprehensive set of motives: "competitiveness", "selfishness", "spite" and "jealousy" (e.g., Barclay, 2006; Kiyonari & Barclay, 2008; Pedersen, et al., 2013; Raihani & Bshary, 2015b,c; Raihani & Bshary, 2019). To look more closely at the reputation of punishers, several trait items covering both possible cooperative and formidable reputational benefits were added as well: "trustworthy", "moral", "cooperative", "dominant", "strong", "competent" and "likeable" (e.g., de Kwaadsteniet et al., 2019; Dhaliwal et al., 2020; Gordon et al., 2014; Gordon & Lea, 2016; Jordan & Kteily, 2020; Jordan & Rand, 2020; Raihani & Bshary, 2015a; Tsai et al., 2022; Wiessner, 2020). Again, all motive and trait items were rated from 1 = "*Not at all*" to 7 = "*Completely*".

For each description actors were additionally asked whether they thought the punitive action reflected more negatively or more positively on them (from 1 = "*Definitely negatively*" to 7 = "*Definitely positively*"; placed at the end of the meta-perception block), whether the punitive action was justified (placed at the end of the self-perception block) and whether they thought the average person would think the punitive action was justified (placed at the end of the meta-perception block). The latter two were rated from 1 = "*Not at all*" to 7 = "*Completely*". An attention check using the same scale instructed participants to select "*completely, the very last button*", to show that they were paying attention (placed at the end of the TPP meta-perception block).

In random order, participants from the first stage of data collection furthermore completed the perspective taking ( $\alpha = 0.85$ ) and empathic-concern ( $\alpha = 0.90$ ) subscales of the Interpersonal Reactivity Index (Davis, 1983), as well as the Big Five Inventory-10 (BFI-10; Rammstedt, 2007; Rammstedt & John, 2007), although I was especially interested in extroversion ( $\alpha = 0.64$ ) and neuroticism ( $\alpha = 0.82$ ) (agreeableness:  $\alpha = 0.43$ ; openness:  $\alpha = 0.58$ ; conscientiousness:  $\alpha = 0.56$ ). All Cronbach's alphas ( $\alpha$ ) reported in this chapter indicate the internal consistency of

each scale based on the data collected in this study. Lastly, they provided demographic information, which completed the survey.

### ***Stage 2 of Data Collection - Observer Participants***

The second stage of data collection took the 294 usable behaviour descriptions provided by 147 actor participants in the first stage of data collection, and presented them as stimuli for observers to rate. Each behaviour description received ratings from between 9 to 12 observers ( $Md = 11$ ,  $M = 10.61$ ,  $SD = 0.51$ ).

Matching the language of the first survey, it was explained that in a first stage of data collection previous participants had been asked to describe recent punishing behaviours they had engaged in, and to predict how others would view them based on each description. Observers were told that they were now invited to take part in the second stage of the study, in which they would be reading eight of those descriptions, based on which they should indicate how they perceive the authors. In contrast to the procedure of Chapter 4, in which observers received the same information around cooperative and uncooperative behaviours as actors did, observer participants now received no further information around punishment and the specific punitive actions actors were asked to describe. The former was done to ensure full transparency for actor and observer participants. Moreover, detailed information and therefore a clearer understanding were believed to lead to more consistent judgments. In contrast, the present study purposefully limited information as to better replicate real-life scenarios, in which people often have to make judgments based on incomplete information. I furthermore aimed to reduce observer bias, as providing information around the two types of punitive actions actors were asked to describe could influence how observers perceive and judge the actors' behaviours. Despite my meta-accuracy studies focusing on different behaviours, the inclusion of both methods across studies could potentially provide insights into the impact of information availability on assessments.

The task prompt for observers to judge behaviour descriptions was identical to that described in Chapter 4, except that it referred to "*punishing behaviours*". Observer participants read and rated either four SPP or four TPP descriptions first, before reading four randomly selected descriptions of the other behaviour. Again, all behaviour descriptions were evenly presented and item order was randomised. The behaviour descriptions provided by actors were not altered in any way, and were presented on the same page as the items to be rated. For each description observers were additionally asked whether they believed the punitive action was justified (from

1= “*Not at all*” to 7 = “*Completely*”) and whether they thought the punitive action reflected more negatively or more positively on the author (from 1 = “*Definitely negatively*” to 7 = “*Definitely positively*”). Using the same scale, observers were asked to select “*definitely positively, the very last button*”, to show that they were paying attention (this attention check was placed in the TPP block).

In contrast to Chapter 4, observer participants additionally completed several measures to assess moderators of accuracy. In random order, observers responded to the BFI-10 (extroversion:  $\alpha = 0.69$ ; neuroticism:  $\alpha = 0.81$ ; agreeableness:  $\alpha = 0.61$ ; openness:  $\alpha = 0.42$ ; conscientiousness:  $\alpha = 0.60$ ; Rammstedt, 2007; Rammstedt & John, 2007), and the perspective taking ( $\alpha = 0.87$ ) and empathic-concern ( $\alpha = 0.91$ ) subscales of the Interpersonal Reactivity Index (Davis, 1983). In addition, I assessed right-wing authoritarianism ( $\alpha = 0.85$ ; six items on a 9-point scale from 1 = “*very strongly disagree*” to 9 = “*very strongly agree*”; Bizumic & Duckitt, 2018), social dominance orientation ( $\alpha = 0.92$ ; eight items on a 7-point scale from 1 = “*Strongly Oppose*” to 7 = “*Strongly Favour*”; Ho et al., 2015) and a left-right political orientation slider with scores from 0 (the far left) to 100 (the far right) with 50 indicating more moderate or centre placements ( $M = 38.55$ ,  $SD = 30.91$ ).

Subsequently, observers were presented with another attention check which asked what topic the previous descriptions had been on. The correct answer was “*punitive actions*”, whilst the incorrect options were “*consumer behaviours*”, “*dietary preferences*” and “*scary stories*”. Finally, they provided demographic information.

#### **5.2.2.4 Analysis Plan**

The analyses conducted for this study are highly similar to those described in the preceding Chapter.

As outlined in section 4.2.4, drawing on the SAM I again used mixed effects modelling to assess baseline accuracy, normative accuracy, distinctive accuracy, transparency bias, meta-insight and trait moderators of accuracy for both observer attributions and actor meta-perceptive accuracy, as well as systematic biases in meta-perceptive accuracy. The details, such as model fitting and centring methods, remain the same.

One deviation is which moderators on meta-perceptive and observer accuracy were examined. The potential moderators on both meta-perceptive and observer accuracy were punishing type,

item, subjective social status, social perspective taking, empathy quotient, and the big five personality traits. Concern for reputation was examined as an additional moderator on meta-perceptive accuracy. Continuous variables were mean-centred.

In addition, to assess which motives are differentially associated with TPP versus SPP for meta-perceptions, self-perceptions and observer motive attributions, I fit three models predicting each of those mean-centred ratings as a function of item, punishing type and their interaction. Subsequently, I performed pairwise comparisons on the estimated marginal means (“emmeans” package; Lenth, 2023) for the interaction between item and punishing type, while adjusting for multiple comparisons with a Bonferroni correction.

To assess whether political ideology influences punitive attitudes, I ran six Cumulative Link Mixed Models (CLMMs) using the ordinal package (Christensen, 2023), fitted with the Laplace approximation. CLMMs are a type of regression model suited for analysing ordinal response data with random effects, which was appropriate given the repeated Likert scale responses used in this study. Hereby, (i) how positively observers believed the punitive action reflected on the punisher, and (ii) how justified they believed the punitive action to be acted as response variable, predicted by either mean-centred (i) social economic conservatism, (ii) RWA, or (iii) SDO. Random intercepts and slopes were included for observer to account for repeated measures within individuals.

Lastly, I fit maximal structure mixed effects models predicting mean-centred approval as a function of mean-centred justness and punishing type, as well as their interaction. This was done separately for observer and actor values.

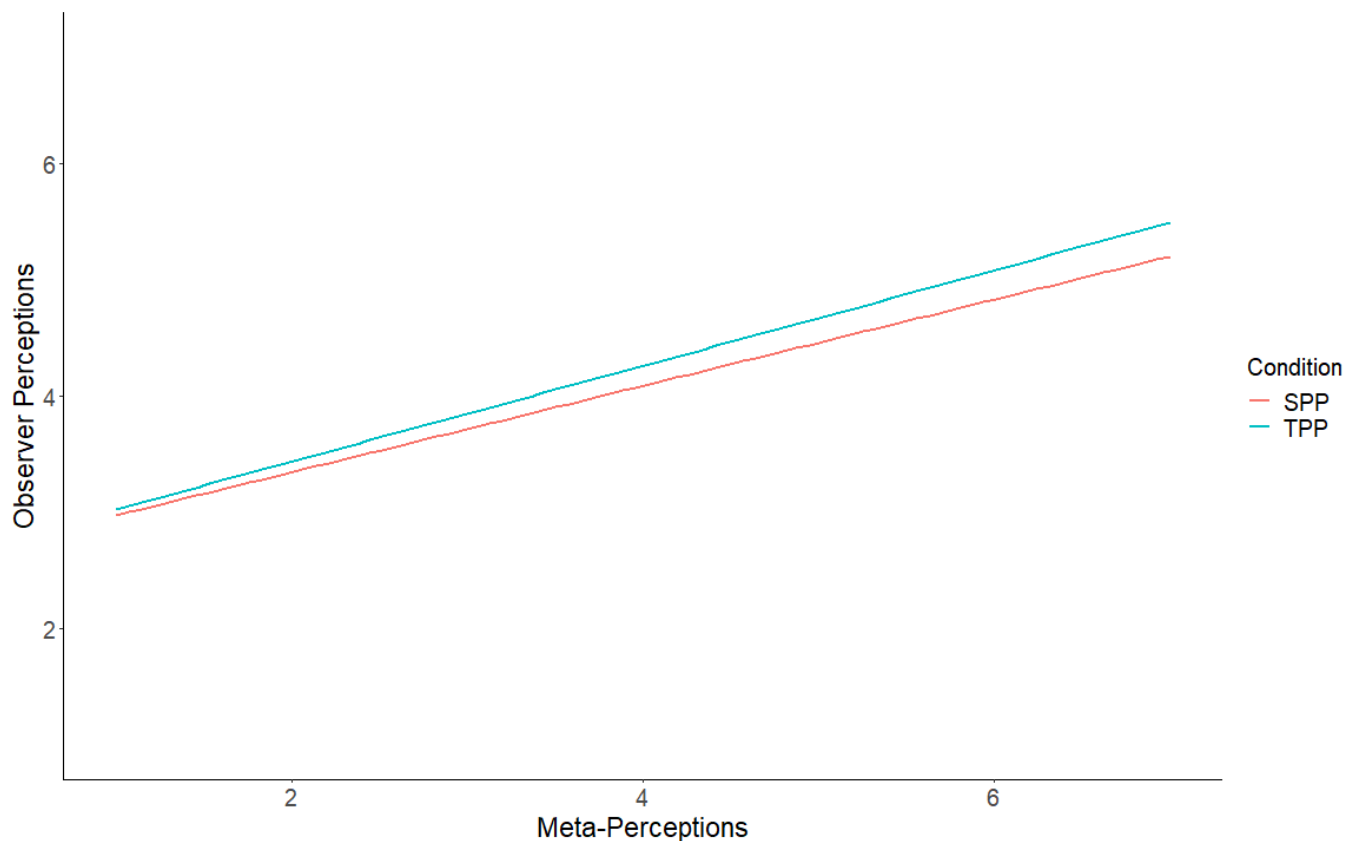
## 5.2.3 Results

### 5.2.3.1 Actor Meta-Accuracy

In simple terms, meta-accuracy refers to the alignment between meta-perceptions (how actors expect to be viewed) and observer perceptions (how observers actually view actors). Figure 5.3 illustrates this alignment, showing that actors' expectations of how they would be perceived closely matched actual observer perceptions, with a particularly strong correspondence for TPP behaviours. Correspondingly, the baseline models found evidence for meta-accuracy in TPP ( $b = 0.40$ , 95% CI [0.37, 0.43],  $t(145.93) = 27.39$ ,  $p < .0001$ ) and SPP ( $b = 0.34$ , 95% CI [0.32, 0.37],  $t(145.85) = 26.59$ ,  $p < .0001$ ) contexts.

**Figure 5.3**

*Relationship between actor meta-perceptions and observer perceptions for TPP and SPP behaviours*



*Note.* The figure presents the raw, uncentered data showing the relationship between actor meta-perceptions (how actors thought others would perceive them) and equivalent observer-perceptions (how observers actually perceived actors) for both TPP and SPP behaviours.

As baseline models conflate two sources of meta-perceptive accuracy – distinctive accuracy (knowledge of how one is uniquely perceived) and normative (accuracy knowledge of how the average actor is perceived), these components were disaggregated in the distinctive accuracy models (see Table 5.2 (TPP behaviours) and Table 5.3 (SPP behaviours) for details on the three main models fit for each behaviour: baseline accuracy models, distinctive accuracy models, and insight models). The distinctive accuracy models found that punishers exhibited meta-perceptive accuracy in predicting how they would be uniquely perceived relative to the average second- ( $b = 0.13$ , 95% CI [0.10, 0.15],  $t(146.89) = 10.73$ ,  $p < .0001$ ) or third-party punisher ( $b = 0.10$ , 95% CI [0.09, 0.12],  $t(144.14) = 12.80$ ,  $p < .0001$ ). The normative profile was also associated with meta-perceptions for both SPP ( $b = 0.80$ , 95% CI [0.73, 0.88],  $t(146.23) = 20.79$ ,  $p < .0001$ ) and TPP ( $b = 0.82$ , 95% CI [0.76, 0.88],  $t(146.31) = 26.85$ ,  $p < .0001$ ). The meta-insight models, which controlled for punishers' true motives and self-perceived traits, found evidence of meta-insight, which means that actors could accurately predict how perceptions of their TPP ( $b = 0.12$ , 95% CI [0.10, 0.13],  $t(145.54) = 13.61$ ,  $p < .0001$ ) and SPP ( $b = 0.09$ , 95% CI [0.07, 0.11],  $t(146.15) = 11.61$ ,  $p < .0001$ ) motives differed from their true motives and self-perceived traits (i.e., they had knowledge of how they would be misperceived). Simultaneously, actors exhibited a transparency bias in both TPP ( $b = 0.67$ , 95% CI [0.64, 0.70],  $t(142.84) = 43.49$ ,  $p < .0001$ ) and SPP ( $b = 0.67$ , 95% CI [0.63, 0.71],  $t(140.10) = 35.74$ ,  $p < .0001$ ) contexts, showing the extent to which they overestimated how accurately observers would perceive them.

**Table 5.2**

*Actor Meta-Perceptive Accuracy for TPP Behaviours Across Models*

| <i>Predictors</i>     | <b>Base Accuracy</b>    |          |                  | <b>Distinctive Accuracy</b> |          |                  | <b>Insight</b>          |          |                  |
|-----------------------|-------------------------|----------|------------------|-----------------------------|----------|------------------|-------------------------|----------|------------------|
|                       | <i>b (CI)</i>           | <i>t</i> | <i>p</i>         | <i>b (CI)</i>               | <i>t</i> | <i>p</i>         | <i>b (CI)</i>           | <i>t</i> | <i>p</i>         |
| (Intercept)           | -0.01<br>(-0.22 – 0.20) | -0.13    | .90              | 0.00<br>(-0.17 – 0.18)      | 0.03     | .98              | -0.00<br>(-0.17 – 0.17) | -0.00    | 0.99             |
| Observer<br>Judgments | 0.40<br>(0.37 – 0.43)   | 27.39    | <<br><b>.001</b> | 0.10<br>(0.09 – 0.12)       | 12.80    | <<br><b>.001</b> | 0.12<br>(0.10 – 0.13)   | 13.61    | <<br><b>.001</b> |
| Normative<br>Profile  |                         |          |                  | 0.82<br>(0.76 – 0.88)       | 26.85    | <<br><b>.001</b> |                         |          |                  |

|                  |                       |       |                  |
|------------------|-----------------------|-------|------------------|
| Self-Perceptions | 0.67<br>(0.64 – 0.70) | 43.49 | <<br><b>.001</b> |
|------------------|-----------------------|-------|------------------|

*Note.* This table presents actor meta-perceptive accuracy for TPP behaviours across the three main model types fit in this chapter. Estimates (*b*), 95% confidence intervals (CIs), *t*-values, and *p*-values are reported. Significant results are bolded. The **baseline accuracy model** tested for a linear relationship between meta-perceptions and observer-perceptions. As meta-accuracy depends both on an understanding of how one is uniquely perceived (*distinctive meta-accuracy*) and an understanding of how the average actor is perceived (*normative accuracy*), the **distinctive accuracy model** tested for a linear relationship between meta-perceptions and the normative profile (the grand mean centred distribution of the self-perception means across actors) on top of the baseline models' linear relationship between meta-perceptions and observer-perceptions. Here, the value for observer judgments reflects distinctive meta-accuracy, while the value for the normative profile reflects normative accuracy. The **insight model** tested for a linear relationship between meta-perceptions and observer-perceptions while controlling for actors' self-perceptions. As this third model controlled for self-perceptions, the relationship between meta-perceptions and observer perceptions can be interpreted as actors' knowledge of how they will be misperceived (*insight*), while the relationship between meta-perception and self-perception can be interpreted as *transparency bias* – the extent to which actors overestimate how accurately they will be perceived. Additionally, as I used person centring (cluster centring within actor) so that meta- and observer-perceptions were anchored to the mean validation criterion within each actor (i.e., meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor), a value of zero represents the average point estimate accuracy within actor, and deviations from zero provide insights into average point-estimate inaccuracies.

**Table 5.3**

*Actor Meta-Perceptive Accuracy for SPP Behaviours Across Models*

| <i>Predictors</i> | <b>Base Accuracy</b> |          |          | <b>Distinctive Accuracy</b> |          |          | <b>Insight</b> |          |          |
|-------------------|----------------------|----------|----------|-----------------------------|----------|----------|----------------|----------|----------|
|                   | <i>b (CI)</i>        | <i>t</i> | <i>p</i> | <i>b (CI)</i>               | <i>t</i> | <i>p</i> | <i>b (CI)</i>  | <i>t</i> | <i>p</i> |



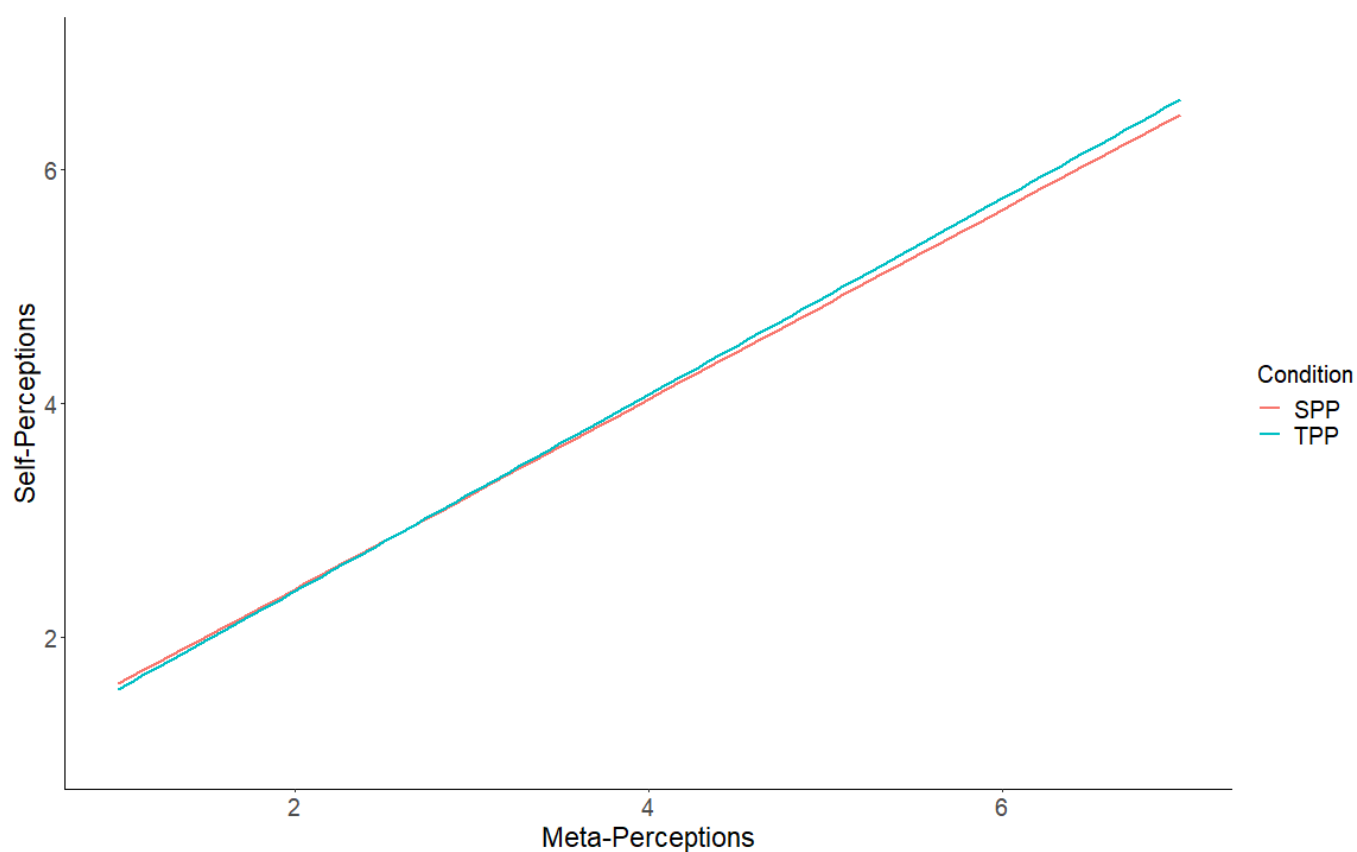
|                       |                         |       |                  |                         |       |                  |                         |       |                  |
|-----------------------|-------------------------|-------|------------------|-------------------------|-------|------------------|-------------------------|-------|------------------|
| (Intercept)           | -0.02<br>(-0.21 – 0.18) | -0.20 | .84              | -0.03<br>(-0.20 – 0.13) | -0.41 | .68              | -0.03<br>(-0.19 – 0.13) | -0.34 | 0.73             |
| Observer<br>Judgments | 0.34<br>(0.32 – 0.37)   | 26.59 | <<br><b>.001</b> | 0.13<br>(0.10 – 0.15)   | 10.74 | <<br><b>.001</b> | 0.09<br>(0.07 – 0.11)   | 11.61 | <<br><b>.001</b> |
| Normative<br>Profile  |                         |       |                  | 0.80<br>(0.73 – 0.88)   | 20.79 | <<br><b>.001</b> |                         |       |                  |
| Self-<br>Perceptions  |                         |       |                  |                         |       |                  | 0.67<br>(0.63 – 0.71)   | 35.74 | <<br><b>.001</b> |

*Note.* This table presents actor meta-perceptive accuracy for SPP behaviours across the three main model types fit in this chapter. Estimates (*b*), 95% confidence intervals (CIs), *t*-values, and *p*-values are reported. Significant results are bolded. The **baseline accuracy model** tested for a linear relationship between meta-perceptions and observer-perceptions. As meta-accuracy depends both on an understanding of how one is uniquely perceived (*distinctive meta-accuracy*) and an understanding of how the average actor is perceived (*normative accuracy*), the **distinctive accuracy model** tested for a linear relationship between meta-perceptions and the normative profile (the grand mean centred distribution of the self-perception means across actors) on top of the baseline models' linear relationship between meta-perceptions and observer-perceptions. Here, the value for observer judgments reflects distinctive meta-accuracy, while the value for the normative profile reflects normative accuracy. The **insight model** tested for a linear relationship between meta-perceptions and observer-perceptions while controlling for actors' self-perceptions. As this third model controlled for self-perceptions, the relationship between meta-perceptions and observer perceptions can be interpreted as actors' knowledge of how they will be misperceived (*insight*), while the relationship between meta-perception and self-perception can be interpreted as *transparency bias* – the extent to which actors overestimate how accurately they will be perceived. Additionally, as I used person centring (cluster centring within actor) so that meta- and observer-perceptions were anchored to the mean validation criterion within each actor (i.e., meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor), a value of zero represents the average point estimate accuracy within actor, and deviations from zero provide insights into average point-estimate inaccuracies.

In line with this bias to overestimate how transparent their attributes are, bivariate correlations between punishers' self- and meta-perceptions across the 147 SPP ( $r(46789) = 0.78$ , 95% CI [0.78, 0.79],  $p < .0001$ ) and 147 TPP behaviour descriptions ( $r(46798) = 0.80$ , 95% CI [0.80, 0.81],  $p < .0001$ ) revealed high correlations (reflected in Figure 5.4). This suggests that punishers did not expect observers' perceptions to deviate significantly from their own.

**Figure 5.4**

*Relationship between actors' meta-perceptions and actors' self-perceptions for TPP and SPP behaviours*



*Note.* The figure presents the raw, uncentered data showing the relationship between actor meta-perceptions (how actors thought others would perceive them) and equivalent actor self-perceptions (how actors viewed themselves) for both TPP and SPP behaviours.

### 5.2.3.2 Moderators of Meta-Accuracy

To test potential moderators of actor meta-accuracy, I separately entered each variable into the distinctive accuracy model as an interaction with observer judgements in predicting meta-perceptions (and as a control with the normative profile as well). A significant interaction would indicate that the trait moderates meta-accuracy, with a positive coefficient meaning actors higher on the trait exhibit significantly higher levels of meta-perceptive accuracy.

Most personality traits did not moderate meta-accuracy: agreeableness did not moderate meta-accuracy for TPP ( $b = 0.002$ , 95% CI  $[-0.01, 0.01]$ ,  $t(144.60) = 0.59$ ,  $p = 0.56$ ) or SPP actions ( $b = -0.006$ , 95% CI  $[-0.02, 0.01]$ ,  $t(144.35) = -0.94$ ,  $p = .35$ ), nor did conscientiousness (TPP:  $b = 0.002$ , 95% CI  $[-0.01, 0.01]$ ,  $t(141.21) = 0.41$ ,  $p = .68$ ; SPP:  $b = 0.001$ , 95% CI  $[-0.01, 0.01]$ ,  $t(144.80) = 0.15$ ,  $p = .88$ ), neuroticism (TPP:  $b = -0.002$ , 95% CI  $[-0.01, 0.00]$ ,  $t(144.50) = -0.63$ ,  $p = .53$ ; SPP:  $b = 0.002$ , 95% CI  $[-0.01, 0.01]$ ,  $t(145.03) = 0.49$ ,  $p = .63$ ) or openness (TPP:  $b = -0.004$ , 95% CI  $[-0.01, 0.00]$ ,  $t(141.64) = -0.95$ ,  $p = .35$ ; SPP:  $b = -0.005$ , 95% CI  $[-0.02, 0.01]$ ,  $t(146.52) = -0.83$ ,  $p = .41$ ). While extroversion also did not moderate meta accuracy for TPP behaviours ( $b = -0.002$ , 95% CI  $[-0.01, 0.01]$ ,  $t(142.09) = -0.54$ ,  $p = .59$ ), more extroverted individuals displayed lower levels of meta-accuracy in second-party punishing contexts ( $b = -0.01$ , 95% CI  $[-0.02, -0.00]$ ,  $t(145.15) p = .03$ ).

For TPP empathic concern did not significantly influence meta-accuracy ( $b = -0.002$ , 95% CI  $[-0.00, 0.00]$ ,  $t(138.42) = -1.57$ ,  $p = .12$ ), but it was significantly linked to the normative profile ( $b = 0.01$ , 95% CI  $[0.00, 0.02]$ ,  $t(143.22) = 2.13$ ,  $p = .03$ ). For SPP empathic concern significantly influenced both meta-accuracy ( $b = -0.005$ , 95% CI  $[-0.01, -0.00]$ ,  $t(142.23) = -2.29$ ,  $p = .02$ ) and normative accuracy ( $b = 0.01$ , 95% CI  $[0.00, 0.03]$ ,  $t(142.83) = 2.29$ ,  $p = .02$ ). Although distinctive meta-accuracy decreased with higher empathic concern, normative accuracy was positively linked to empathic concern. Perspective taking also did not significantly influence meta-accuracy for TPP ( $b = -0.0008$ , 95% CI  $[-0.00, 0.00]$ ,  $t(140.92) = -0.48$ ,  $p = .63$ ) or SPP ( $b = -0.004$ , 95% CI  $[-0.00, 0.00]$ ,  $t(144.87) = -1.76$ ,  $p = .08$ ), although it did improve normative accuracy for both third-party ( $b = 0.02$ , 95% CI  $[0.01, 0.03]$ ,  $t(144.65) = 3.47$ ,  $p = .0007$ ) and second-party punishers ( $b = 0.03$ , 95% CI  $[0.01, 0.03]$ ,  $t(144.93) = 3.78$ ,  $p = .0002$ ).

Lastly, punishing type significantly moderated distinctive meta-accuracy ( $b = -0.03$ , 95% CI  $[-0.04, -0.02]$ ,  $t(93047.53) = -5.70$ ,  $p < .0001$ ) and normative accuracy ( $b = 0.03$ , 95% CI  $[0.01,$

0.04],  $t(93196.26) = 2.90$ ,  $p = .004$ ), in that punishers displayed significantly higher levels of distinctive accuracy in SPP contexts, but higher normative accuracy in TPP contexts.

As longer, more detailed responses could lead to higher accuracy, average word length for SPP and TPP behaviour descriptions were additionally compared. However, actors provided similar-length descriptions for both TPP ( $M = 59.76$ ,  $SD = 47.47$ , range = 8 – 405 words) and SPP ( $M = 62.11$ ,  $SD = 43.99$ , range = 11-307 words),  $t(290.32) = -0.44$ , 95% CI [-12.85, 8.16],  $p = .66$ .

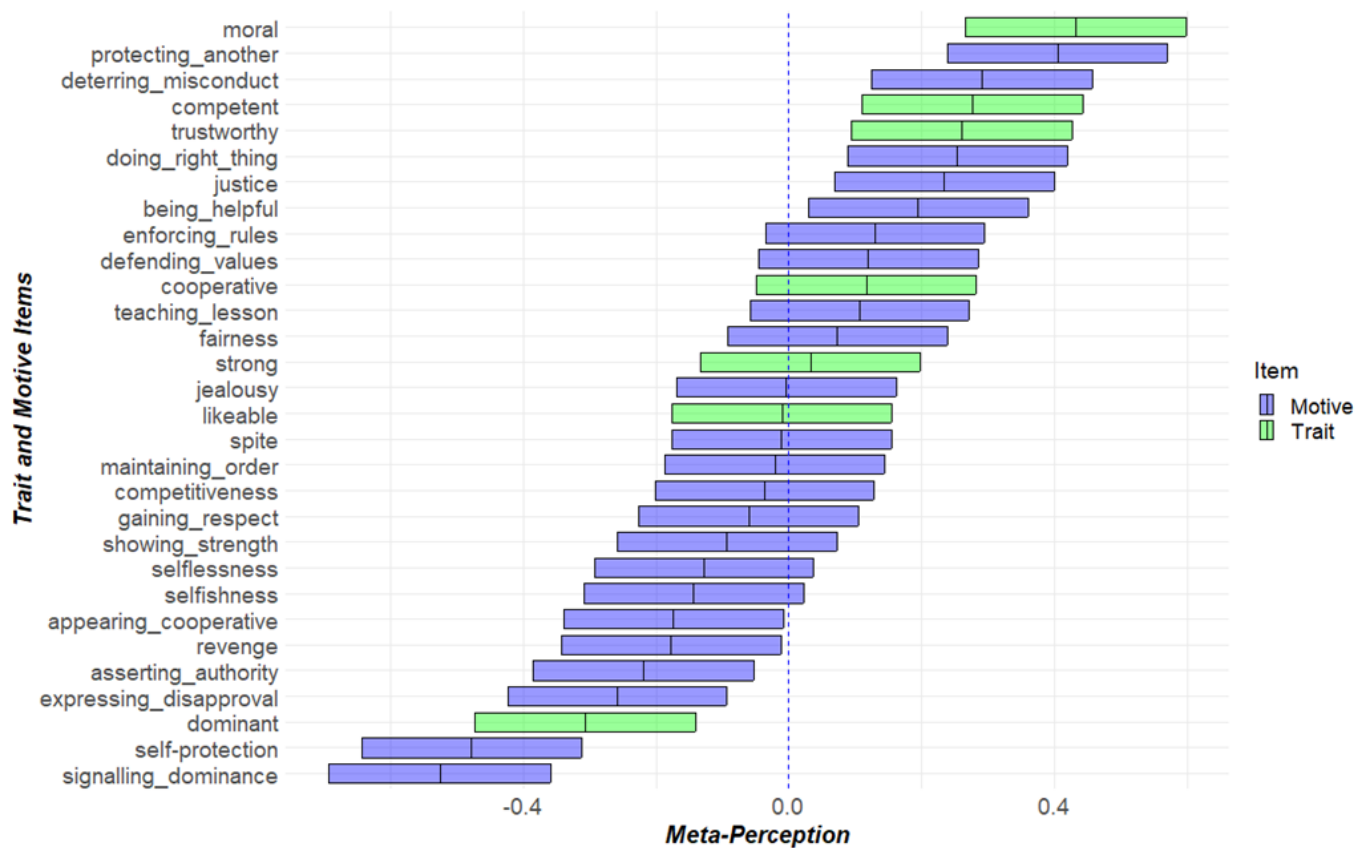
### 5.2.3.3 Meta-Bias

To investigate systematic biases in meta-perceptions I variable-centred the data and incorporated the item variable into the baseline meta-accuracy models, interacting item with meta-accuracy, before calculating marginal means. By variable-centring the data, estimates that deviate significantly from zero can be understood as systematic directional bias within a particular perception.

When predicting observer evaluations of third-party punishing actions, actors overestimated how moral, competent, and trustworthy they would be perceived (see Figure 5.5). Third-party punishers also overestimated how much observers would think their punitive actions were motivated by a desire to protect someone else, to do the right thing, to bring justice, or to be helpful. Simultaneously, they underestimated how much observers would think their TPP action would be viewed as an attempt to signal dominance, protect themselves, express disapproval, assert authority or be out of revenge. They also underestimated how dominant observers would perceive them. Overall, this suggests the presence of a positive meta-bias.

### Figure 5.5

*Visualisation of systematic meta-biases in third-party punishment using estimated marginal means*

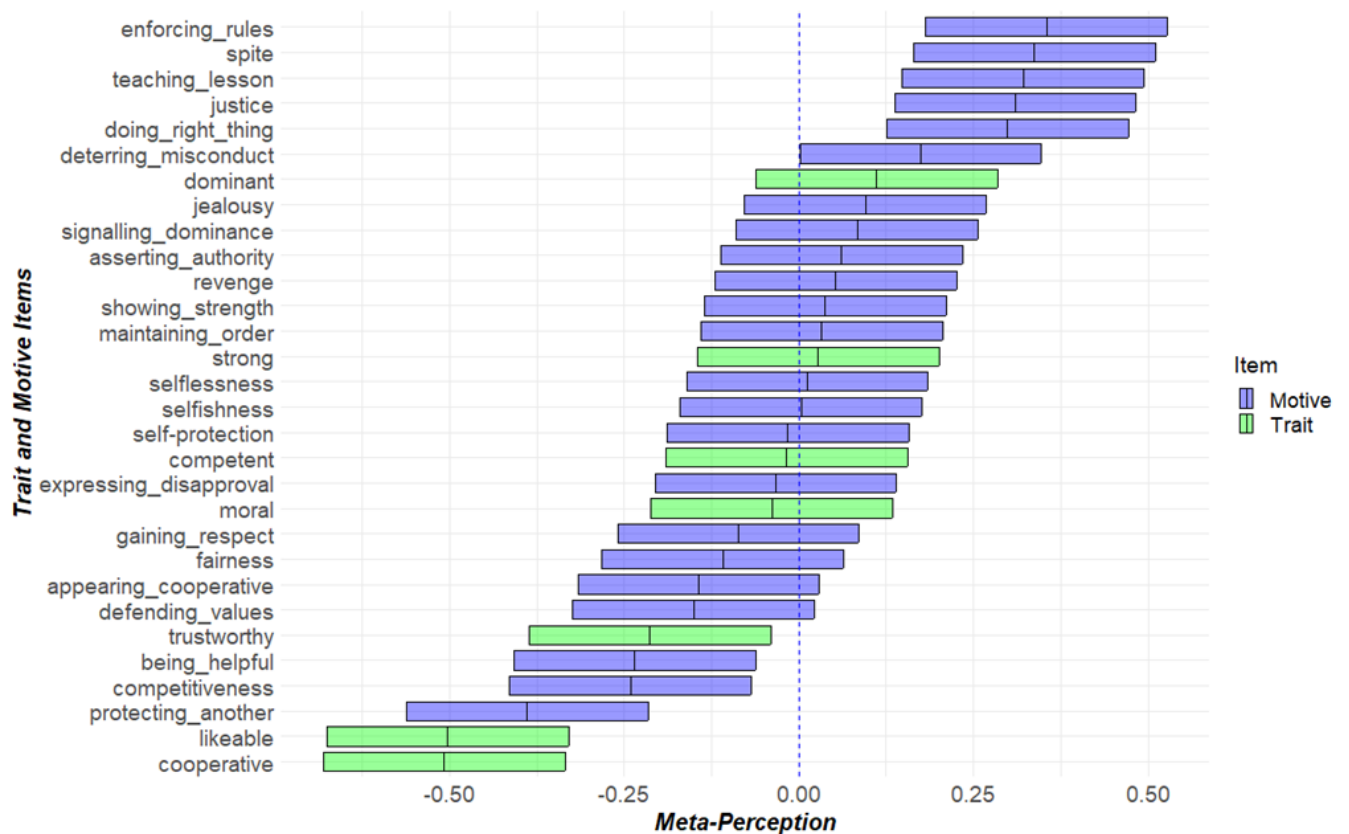


*Note.* This figure displays unstandardised estimated marginal means of actor meta-perceptions for various traits and motives, represented by the horizontal lines within each crossbar. The bars denote 95% CIs. The dashed blue line at zero represents mean-level accuracy, meaning that estimates which deviate from zero can be interpreted as directional bias (i.e., items to the left represent underestimation whilst items to the right of the line represent overestimation).

Conversely, actors overestimated how much observers would believe their SPP actions were motivated by spite, justice, a desire to enforce rules, teach a lesson, and do the right thing (see Figure 5.6). They also underestimated how cooperative, likeable and trustworthy observers would think they were, as well as underestimating SPP motive attributions of protecting others, competitiveness, and being helpful. Therefore, there is a less clear valence bias for SPP situations.

**Figure 5.6**

*Visualisation of systematic meta-biases in second-party punishment using estimated marginal means*



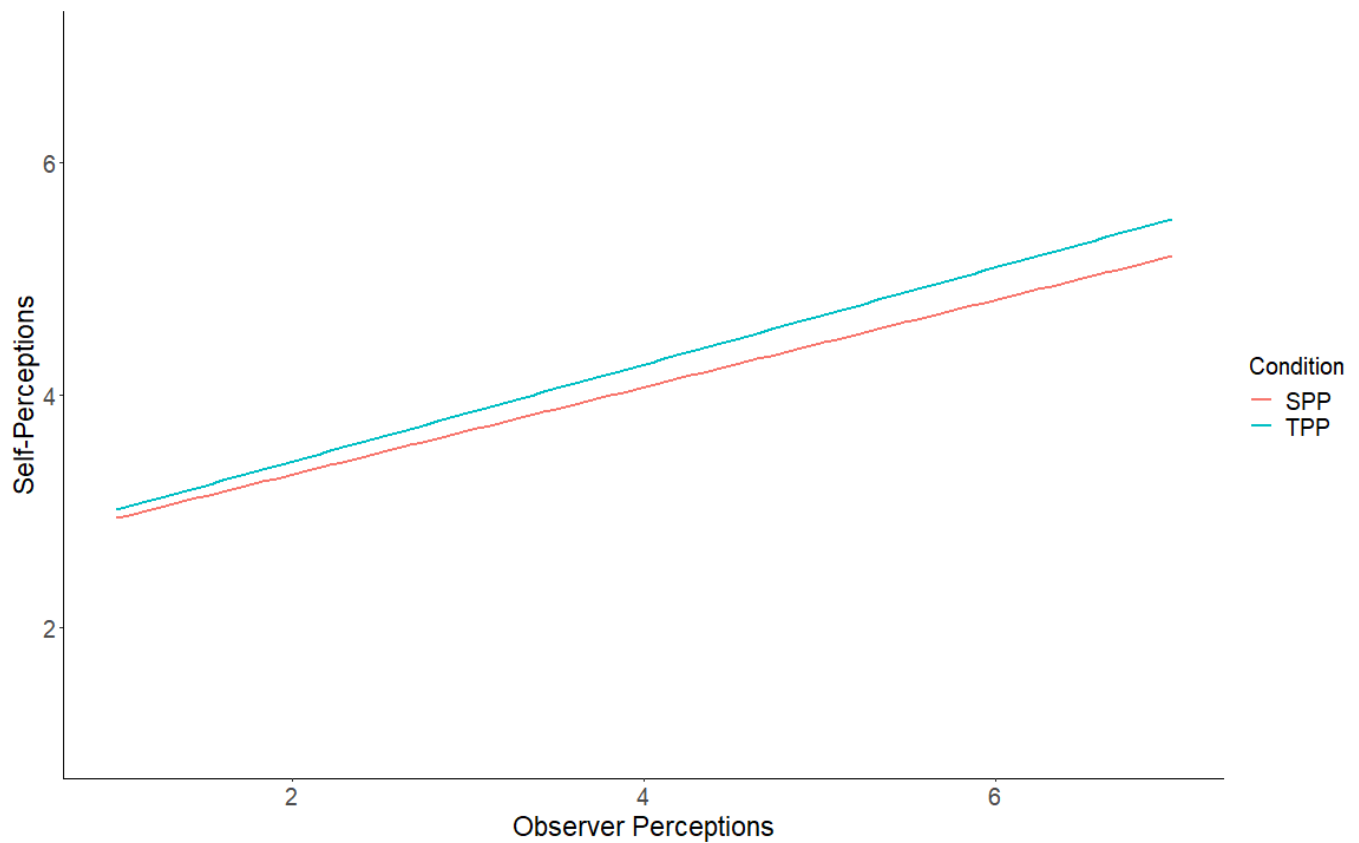
*Note.* This figure displays unstandardised estimated marginal means of actor meta-perceptions for various traits and motives, represented by the horizontal lines within each crossbar. The bars denote 95% CIs. The dashed blue line at zero represents mean-level accuracy, meaning that estimates which deviate from zero can be interpreted as directional bias (i.e., items to the left represent underestimation whilst items to the right of the line represent overestimation).

### 5.2.3.4 Observer Accuracy

In simple terms, observer accuracy is determined by the alignment between observer perceptions (how observers perceive actors) and actors' self-perceptions (how actors perceive themselves, which serves as a proxy for their true motives and traits). Figure 5.7 illustrates this alignment, showing that observers' evaluations of actors closely matched the motives and traits reported by the actors.

**Figure 5.7**

*Relationship between observer perceptions and actor self-perceptions for TPP and SPP behaviours*



*Note.* The figure presents the raw, uncentered data showing the relationship between observer-perceptions (how observers perceived actors) and equivalent actor self-perceptions (how actors perceived themselves) for both TPP and SPP behaviours.

Correspondingly, the baseline models found that observers were able to accurately assess second- ( $b = 0.44$ , 95% CI [0.41, 0.48],  $t(179.87) = 26.31$ ,  $p < .0001$ ) and third-party punishers' ( $b = 0.49$ , 95% CI [0.46, 0.53],  $t(211.39) = 29.15$ ,  $p < .0001$ ) actual motives and trait self-perceptions. As observer accuracy may be influenced by observers' reliance on preconceived notions or general knowledge about common motives and traits in the population, rather than focusing solely on the specific motives and traits displayed in the behaviour they are evaluating, caution is required when interpreting baseline models. Thus, distinctive accuracy models controlled for the average distribution of motives and traits by examining how well estimated values (observer perceptions) aligned with actual values (self-reported true motives and traits), while also assessing the alignment of estimated values with the normative profile. I found evidence for distinctive accuracy (TPP:  $b = 0.14$ , 95% CI [0.11, 0.16],  $t(172.39) = 10.22$ ,  $p < .0001$ ; SPP:  $b = 0.18$ , 95% CI [0.15, 0.21],  $t(174.92) = 11.55$ ,  $p < .0001$ ), and normative accuracy (TPP:  $b = 0.70$ , 95% CI [0.64, 0.76],  $t(240.13) = 22.13$ ,  $p < .0001$ ; SPP:  $b = 0.70$ , 95% CI [0.64, 0.76],  $t(231.93) = 23.22$ ,  $p < .0001$ ). Furthermore, they exhibited observer insight by

accurately assessing actors' true motives and traits beyond how second- ( $b = 0.27$ , 95% CI [0.23, 0.31],  $t(163.58) = 13.44$ ,  $p < .0001$ ) and third-party punishers ( $b = 0.28$ , 95% CI [0.24, 0.32],  $t(162.01) = 14.27$ ,  $p < .0001$ ) expected to be perceived. At the same time, the observer insight model found evidence for opacity (TPP:  $b = 0.29$ , 95% CI [0.26, 0.33],  $t(159.67) = 14.79$ ,  $p < .0001$ ; SPP:  $b = 0.25$ , 95% CI [0.21, 0.29],  $t(174.23) = 12.37$ ,  $p < .0001$ ), meaning that observers' judgments were associated with punishers' meta-perceptions above punishers' self-perceptions. See Table 5.4 (TPP actions) and Table 5.5 (SPP actions) for details on the three main models fit for each behaviour: baseline accuracy models, distinctive accuracy models, and insight models.

**Table 5.4**

*Observer Judgement Accuracy for TPP Behaviours Across Models*

| <i>Predictors</i> | <b>Base Accuracy</b>    |          |                  | <b>Distinctive Accuracy</b> |          |                  | <b>Insight</b>          |          |                  |
|-------------------|-------------------------|----------|------------------|-----------------------------|----------|------------------|-------------------------|----------|------------------|
|                   | <i>b (CI)</i>           | <i>t</i> | <i>p</i>         | <i>b (CI)</i>               | <i>t</i> | <i>p</i>         | <i>b (CI)</i>           | <i>t</i> | <i>p</i>         |
| (Intercept)       | -0.00<br>(-0.17 – 0.16) | -0.02    | .98              | -0.00<br>(-0.17 – 0.17)     | -0.02    | .99              | -0.04<br>(-0.24 – 0.17) | -0.35    | .73              |
| Self-Perceptions  | 0.49<br>(0.46 – 0.53)   | 29.15    | <b>&lt; .001</b> | 0.14<br>(0.11 – 0.16)       | 10.22    | <b>&lt; .001</b> | 0.28<br>(0.24 – 0.32)   | 14.27    | <b>&lt; .001</b> |
| Normative Profile |                         |          |                  | 0.70<br>(0.64 – 0.76)       | 22.13    | <b>&lt; .001</b> |                         |          |                  |
| Meta-Perceptions  |                         |          |                  |                             |          |                  | 0.29<br>(0.25 – 0.33)   | 14.79    | <b>&lt; .001</b> |

*Note.* This table presents observer judgement accuracy for TPP behaviours across the three main model types fit in this chapter. Estimates ( $b$ ), 95% confidence intervals (CIs),  $t$ -values, and  $p$ -values are reported. Significant results are bolded. The **baseline accuracy model** tested for a linear relationship between observer perceptions and actor self-perceptions. As observer accuracy depends both on an understanding of an actors' unique traits and motives relative to other actors (*distinctive meta-accuracy*) and an understanding of which motives and traits are common in the population (*normative accuracy*), the distinctive accuracy model tested for a linear relationship between observer-perceptions and the normative profile (the grand mean



centred distribution of the self-perception means across actors) on top of the baseline models' linear relationship between observer-perceptions and actor self-perceptions. Here, the value for self-perceptions reflects distinctive accuracy, while the value for the normative profile reflects normative accuracy. The *insight model* tested for a linear relationship between observer-perceptions and actor self-perceptions (*insight*) while controlling for actors' meta-perceptions (*opaqueness*). Observer insight reflects the observer's ability to accurately perceive the disparity between an actor's actual motives and traits and the actor's predicted perceptions. Conversely, opaqueness is exhibited when observers perceptions are associated with actors' meta-perceptions above actors' self-reported actual motives and traits, which would mean that observers perceive some items inaccurately and in a manner that actors anticipated. Additionally, as I used person centring (cluster centring within actor) so that meta- and observer-perceptions were anchored to the mean validation criterion within each actor (i.e., meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor), a value of zero represents the average point estimate accuracy within actor, and deviations from zero provide insights into average point-estimate inaccuracies.

**Table 5.5**

*Observer Judgement Accuracy for SPP Behaviours Across Models*

| <i>Predictors</i> | <b>Base Accuracy</b>   |           |          | <b>Distinctive Accuracy</b> |           |          | <b>Insight</b>          |           |          |
|-------------------|------------------------|-----------|----------|-----------------------------|-----------|----------|-------------------------|-----------|----------|
|                   | <i>b (CI)</i>          | <i>t</i>  | <i>p</i> | <i>b (CI)</i>               | <i>t</i>  | <i>p</i> | <i>b (CI)</i>           | <i>t</i>  | <i>p</i> |
| (Intercept)       | 0.01<br>(-0.16 – 0.18) | 0.08      | .93      | 0.01<br>(-0.16 – 0.18)      | 0.08      | .93      | -0.01<br>(-0.21 – 0.18) | -0.12     | .90      |
| Self-Perceptions  | 0.44<br>(0.41 – 0.48)  | 26.3<br>1 | < .001   | 0.18<br>(0.15 – 0.21)       | 11.5<br>5 | < .001   | 0.27<br>(0.23 – 0.31)   | 13.4<br>4 | < .001   |
| Normative Profile |                        |           |          | 0.70<br>(0.64 – 0.76)       | 23.2<br>2 | < .001   |                         |           |          |

|                  |                           |           |        |
|------------------|---------------------------|-----------|--------|
| Meta-Perceptions | 0.25<br>(0.21 – 0.29<br>) | 12.3<br>7 | < .001 |
|------------------|---------------------------|-----------|--------|

*Note.* This table presents observer judgement accuracy for SPP behaviours across the three main model types fit in this chapter. Estimates (*b*), 95% confidence intervals (CIs), *t*-values, and *p*-values are reported. Significant results are bolded. The **baseline accuracy model** tested for a linear relationship between observer perceptions and actor self-perceptions. As observer accuracy depends both on an understanding of an actors' unique traits and motives relative to other actors (*distinctive meta-accuracy*) and an understanding of which motives and traits are common in the population (*normative accuracy*), the distinctive accuracy model tested for a linear relationship between observer-perceptions and the normative profile (the grand mean centred distribution of the self-perception means across actors) on top of the baseline models' linear relationship between observer-perceptions and actor self-perceptions. Here, the value for self-perceptions reflects distinctive accuracy, while the value for the normative profile reflects normative accuracy. The **insight model** tested for a linear relationship between observer-perceptions and actor self-perceptions (*insight*) while controlling for actors' meta-perceptions (*opaqueness*). Observer insight reflects the observer's ability to accurately perceive the disparity between an actor's actual motives and traits and the actor's predicted perceptions. Conversely, opaqueness is exhibited when observers perceptions are associated with actors' meta-perceptions above actors' self-reported actual motives and traits, which would mean that observers perceive some items inaccurately and in a manner that actors anticipated. Additionally, as I used person centring (cluster centring within actor) so that meta- and observer-perceptions were anchored to the mean validation criterion within each actor (i.e., meta-perceptions were true-mean centred on the average observer judgment towards the specific actor, observer attributions were true-mean centred on the average self-perception provided by the actor individuals, and actors' self-perceptions were mean-centred within actor), a value of zero represents the average point estimate accuracy within actor, and deviations from zero provide insights into average point-estimate inaccuracies.

### 5.2.3.5 Moderators of Observer Accuracy

Observers were significantly more accurate at predicting the motives and traits of second- compared to third-party punishers ( $b = -0.03$ , 95% CI  $[-0.05, -0.02]$ ,  $t(90296.44) = -4.58$ ,  $p < .0001$ ) with no significant difference in normative accuracy ( $b = -0.01$ , 95% CI  $[-0.03, 0.01]$ ,

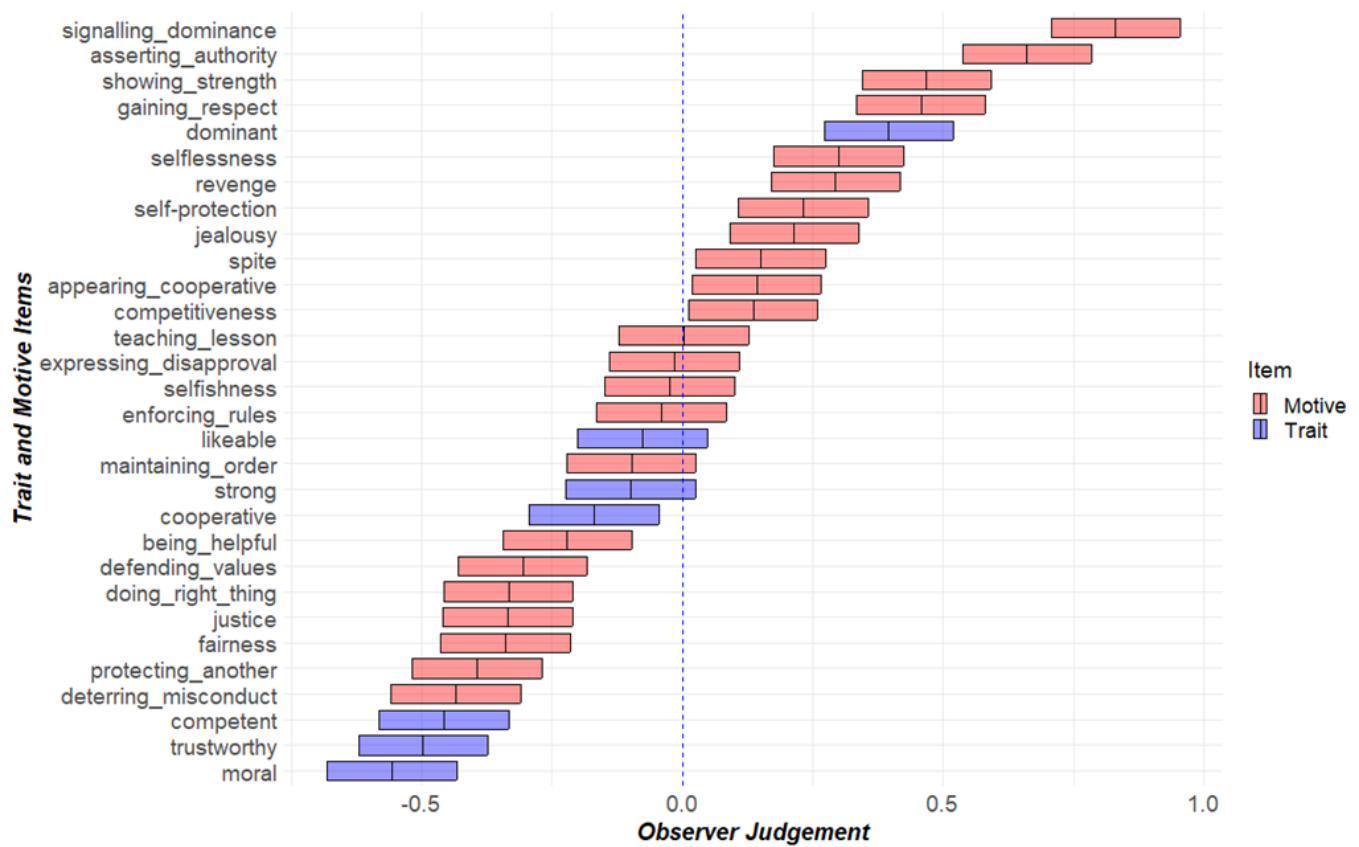
$t(91146.34) = -0.97, p = .33$ ). Neither extroversion (TPP:  $b = 0.0009$ , 95% CI  $[-0.01, 0.01]$ ,  $t(362.28) = 0.31, p = .76$ ; SPP:  $b = 0.0005$ , 95% CI  $[-0.01, 0.01]$ ,  $t(366.61) = 0.15, p = .88$ ), agreeableness (TPP:  $b = -0.002$ , 95% CI  $[-0.01, 0.00]$ ,  $t(363.25) = -0.48, p = .63$ ; SPP:  $b = -0.003$ , 95% CI  $[-0.01, 0.00]$ ,  $t(374.50) = -0.92, p = .36$ ), conscientiousness (TPP:  $b = -0.002$ , 95% CI  $[-0.01, 0.01]$ ,  $t(353.02) = -0.46, p = .65$ ; SPP:  $b = 0.006$ , 95% CI  $[-0.00, 0.01]$ ,  $t(373.81) = 1.62, p = .11$ ), neuroticism (TPP:  $b = 0.001$ , 95% CI  $[-0.00, 0.01]$ ,  $t(372.28) = 0.53, p = .60$ ; SPP:  $b = -0.002$ , 95% CI  $[-0.01, 0.00]$ ,  $t(369.19) = -0.84, p = .40$ ), openness (TPP:  $b = 0.006$ , 95% CI  $[-0.00, 0.01]$ ,  $t(360.76) = 1.73, p = .08$ ; SPP:  $b = 0.0003$ , 95% CI  $[-0.01, 0.01]$ ,  $t(379.83) = 0.08, p = .93$ ), empathic concern ( $b = 0.0003$ , 95% CI  $[-0.00, 0.00]$ ,  $t(352.97) = 0.23, p = .82$ ; SPP:  $b = 0.0007$ , 95% CI  $[-0.00, 0.00]$ ,  $t(375.85) = 0.60, p = .55$ ), nor perspective taking (TPP:  $b = 0.0004$ , 95% CI  $[-0.00, 0.00]$ ,  $t(348.99) = 0.34, p = .74$ ; SPP:  $b = 0.002$ , 95% CI  $[-0.00, 0.00]$ ,  $t(381.53) = 1.18, p = .24$ ) moderated distinctive observer accuracy. However, normative accuracy improved for more extroverted ( $b = -0.02$ , 95% CI  $[-0.03, -0.00]$ ,  $t(389.24) = -2.49, p = .01$ ) and conscientious ( $b = 0.02$ , 95% CI  $[0.01, 0.04]$ ,  $t(388.94) = 2.53, p = .01$ ) observers in TPP contexts, as well as for observers higher in empathic concern (TPP:  $b = 0.01$ , 95% CI  $[0.01, 0.02]$ ,  $t(390.08) = 3.95, p = .0001$ ; SPP:  $b = 0.009$ , 95% CI  $[0.00, 0.01]$ ,  $t(382.18) = 3.34, p = .0009$ ) and perspective taking (TPP:  $b = 0.02$ , 95% CI  $[0.01, 0.02]$ ,  $t(386.43) = 4.95, p < .0001$ ; SPP:  $b = 0.01$ , 95% CI  $[0.00, 0.02]$ ,  $t(391.80) = 3.23, p = .001$ ). More agreeable observers in SPP contexts displayed reduced normative accuracy ( $b = -0.02$ , 95% CI  $[-0.03, -0.00]$ ,  $t(386.47) = -2.09, p = .04$ ).

### 5.2.3.6 Observer Bias

When evaluating third-party punishers, observers overestimated how much actors were motivated by a desire to signal dominance, assert authority, show strength, gain respect, selflessness, revenge, self-protection, jealousy, spite, appear cooperative, and competitiveness (Figure 5.8). Conversely, observers rated third-party punishers as less moral, trustworthy, competent, and cooperative than actors perceived themselves. Additionally, observers underestimated how much punishers were motivated by a desire to deter misconduct, protect another, fairness, justice, doing the right thing, defending values, and being helpful. This pattern indicates a negativity bias.

**Figure 5.8**

*Visualisation of systematic observer biases in third-party punishment using estimated marginal means*



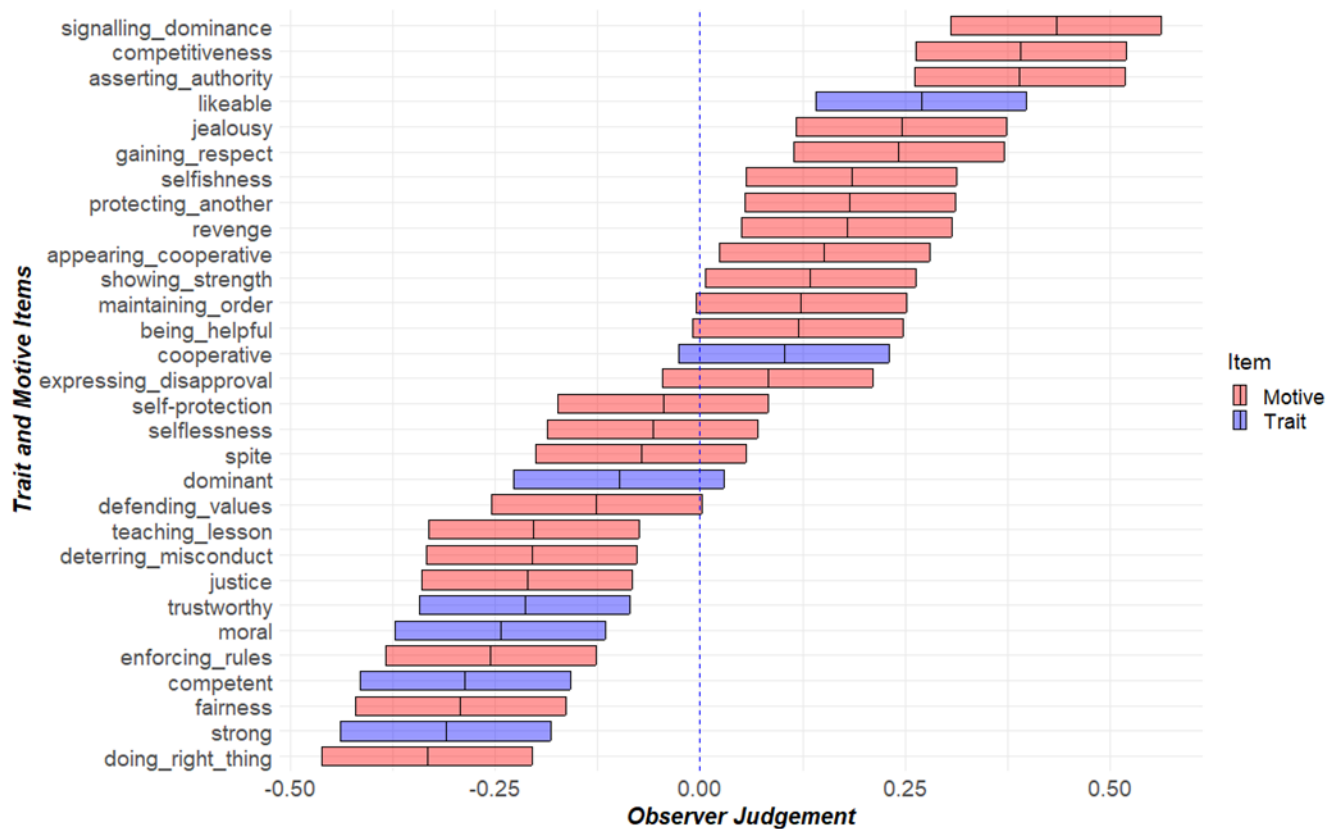
*Note.* This figure displays unstandardised estimated marginal means of observer perceptions for various traits and motives, represented by the horizontal lines within each crossbar. The bars denote 95% CIs. The dashed blue line at zero represents mean-level accuracy, meaning that estimates which deviate from zero can be interpreted as directional bias (i.e., items to the left represent underestimation whilst items to the right of the line represent overestimation).

Similarly, when evaluating second-party punishers, observers overestimated how much actors were motivated by a desire to signal dominance, competitiveness, asserting authority, jealousy, gaining respect, selfishness, protecting others, revenge, appearing cooperative and showing strength, as well as rating second-party punishers as more likeable than punishers rated themselves (Figure 5.9). Moreover, observers perceived punishers as less strong, competent, moral, and trustworthy than second-party punishers perceived themselves to be, and underestimated how much second-party punishers were motivated by doing the right thing, fairness, enforcing rules or regulations, justice, deterring future misconduct, and teaching a

lesson. This again indicates a negativity bias, as observers overestimated mostly formidability and other self-centred concerns (besides also overestimating the desire to protect others and rating them as more likeable), whilst underestimating other-regarding punishing motives and viewing punishers' traits more negatively.

**Figure 5.9**

*Visualisation of systematic observer biases in second-party punishment using estimated marginal means*



*Note.* This figure displays unstandardised estimated marginal means of observer perceptions for various traits and motives, represented by the horizontal lines within each crossbar. The bars denote 95% CIs. The dashed blue line at zero represents mean-level accuracy, meaning that estimates which deviate from zero can be interpreted as directional bias (i.e., items to the left represent underestimation whilst items to the right of the line represent overestimation).

### 5.2.3.7 Item Relevance for SPP versus TPP

To assess which motives are differentially associated with TPP versus SPP for observer perceptions, meta-perceptions, and self-perceptions, I fit three models predicting each of those mean-centred ratings as a function of item, punishing type and their interaction. Subsequently, I performed pairwise comparisons on the estimated marginal means for the interaction between item and punishing type, while adjusting for multiple comparisons with a Bonferroni correction (Table 5.6). To see how the relative motive positions match between observer-, meta-, and self-perceptions see Appendix D7.

**Table 5.6**

*Estimates (b), t-values, and p-values for the applicability of motives and traits to TPP versus SPP behaviours*

|                            | Observer Perception |          |          | Meta-Perception |          |          | Self-Perception |          |          |
|----------------------------|---------------------|----------|----------|-----------------|----------|----------|-----------------|----------|----------|
|                            | <i>b</i>            | <i>t</i> | <i>p</i> | <i>b</i>        | <i>t</i> | <i>p</i> | <i>b</i>        | <i>t</i> | <i>p</i> |
| Protecting another (M)     | -3.18               | -47.58   | < .0001  | -3.88           | -45.50   | < .0001  | -3.75           | -46.88   | < .0001  |
| Self-protection (M)        | 2.01                | 30.07    | < .0001  | 2.38            | 27.98    | < .0001  | 2.29            | 28.59    | < .0001  |
| Being helpful (M)          | -1.71               | -25.56   | < .0001  | -2.09           | -24.49   | < .0001  | -2.05           | -25.55   | < .0001  |
| Selflessness (M)           | -1.32               | -19.72   | < .0001  | -1.18           | -13.86   | < .0001  | -0.96           | -11.96   | < .0001  |
| Doing right thing (M)      | -1.04               | -15.61   | < .0001  | -1.02           | -12.00   | < .0001  | -1.04           | -13.00   | < .0001  |
| Spite (M)                  | 0.90                | 13.45    | < .0001  | 1.18            | 13.88    | < .0001  | 1.12            | 14.04    | < .0001  |
| Maintaining order (M)      | -0.82               | -12.27   | < .0001  | -0.75           | -8.77    | < .0001  | -1.04           | -12.96   | < .0001  |
| Revenge (M)                | 0.81                | 12.19    | < .0001  | 1.06            | 12.41    | < .0001  | 0.93            | 11.64    | < .0001  |
| Selfishness (M)            | 0.75                | 11.27    | < .0001  | 0.93            | 10.93    | < .0001  | 0.55            | 6.82     | < .0001  |
| Appearing cooperative (M)  | -0.71               | -10.62   | < .0001  | -0.67           | -7.86    | < .0001  | -0.72           | -8.94    | < .0001  |
| Defending values (M)       | -0.69               | -10.38   | < .0001  | -0.97           | -11.39   | < .0001  | -0.87           | -10.87   | < .0001  |
| Fairness (M)               | -0.61               | -9.11    | < .0001  | -0.80           | -9.34    | < .0001  | -0.65           | -8.16    | < .0001  |
| Moral (T)                  | -0.58               | -8.70    | < .0001  | -1.05           | -12.36   | < .0001  | -0.89           | -11.15   | < .0001  |
| Cooperative (T)            | -0.56               | -8.38    | < .0001  | -1.16           | -13.61   | < .0001  | -0.83           | -10.35   | < .0001  |
| Trustworthy (T)            | -0.51               | -7.60    | < .0001  | -0.98           | -11.51   | < .0001  | -0.79           | -9.84    | < .0001  |
| Likeable (T)               | -0.50               | -7.44    | < .0001  | -0.98           | -11.51   | < .0001  | -0.84           | -10.50   | < .0001  |
| Justice (M)                | -0.44               | -6.53    | < .0001  | -0.35           | -4.10    | .0001    | -0.56           | -6.96    | < .0001  |
| Enforcing rules (M)        | -0.42               | -6.25    | < .0001  | -0.24           | -2.82    | .005     | -0.20           | -2.49    | .01      |
| Strong (T)                 | -0.36               | -5.43    | < .0001  | -0.39           | -4.54    | < .0001  | -0.15           | -1.87    | .06      |
| Gaining respect (M)        | -0.34               | -5.09    | < .0001  | -0.39           | -4.59    | < .0001  | -0.12           | -1.53    | .13      |
| Competent (T)              | -0.33               | -4.94    | < .0001  | -0.66           | -7.72    | < .0001  | -0.50           | -6.21    | < .0001  |
| Deterring misconduct (M)   | -0.26               | -3.87    | .0001    | -0.37           | -4.31    | < .0001  | -0.48           | -6.06    | < .0001  |
| Competitiveness (M)        | 0.26                | 3.86     | < .0001  | 0.08            | 0.93     | .36      | 0.00            | 0.06     | .95      |
| Dominant (T)               | -0.25               | -3.77    | .0002    | 0.14            | 1.60     | .11      | 0.25            | 3.07     | .002     |
| Jealousy (M)               | 0.21                | 3.16     | .002     | 0.34            | 4.00     | .0001    | 0.18            | 2.28     | .02      |
| Signalling dominance (M)   | 0.19                | 2.84     | .005     | 0.77            | 9.00     | < .0001  | 0.59            | 7.36     | < .0001  |
| Asserting authority (M)    | 0.15                | 2.20     | .03      | 0.40            | 4.67     | < .0001  | 0.42            | 5.26     | < .0001  |
| Expressing disapproval (M) | -0.11               | -1.71    | .09      | 0.12            | 1.42     | .16      | -0.21           | -2.61    | .009     |
| Showing strength (M)       | -0.05               | -0.78    | .44      | 0.08            | 0.90     | .37      | 0.28            | 3.56     | .0004    |
| Teaching lesson (M)        | -0.04               | -0.60    | .55      | 0.18            | 2.06     | .04      | 0.17            | 2.10     | .04      |

*Note.* This table displays estimates (*b*), *t*-values, and *p*-values for the applicability of motives and traits to SPP versus TPP behaviours. A Bonferroni correction was applied to adjust for

multiple comparisons across all rating types (observer perceptions, meta-perceptions, self-perceptions). The adjusted significance level is  $\alpha = .0009$  (i.e.,  $0.05/54$ ). Blue indicates items significantly more applicable to TPP, orange indicates items significantly more applicable to SPP, and grey indicates items with no significant difference. 'M' signifies motive items, and 'T' signifies trait items. Items are ordered based on the magnitude of the *b*-estimate for observer perceptions.

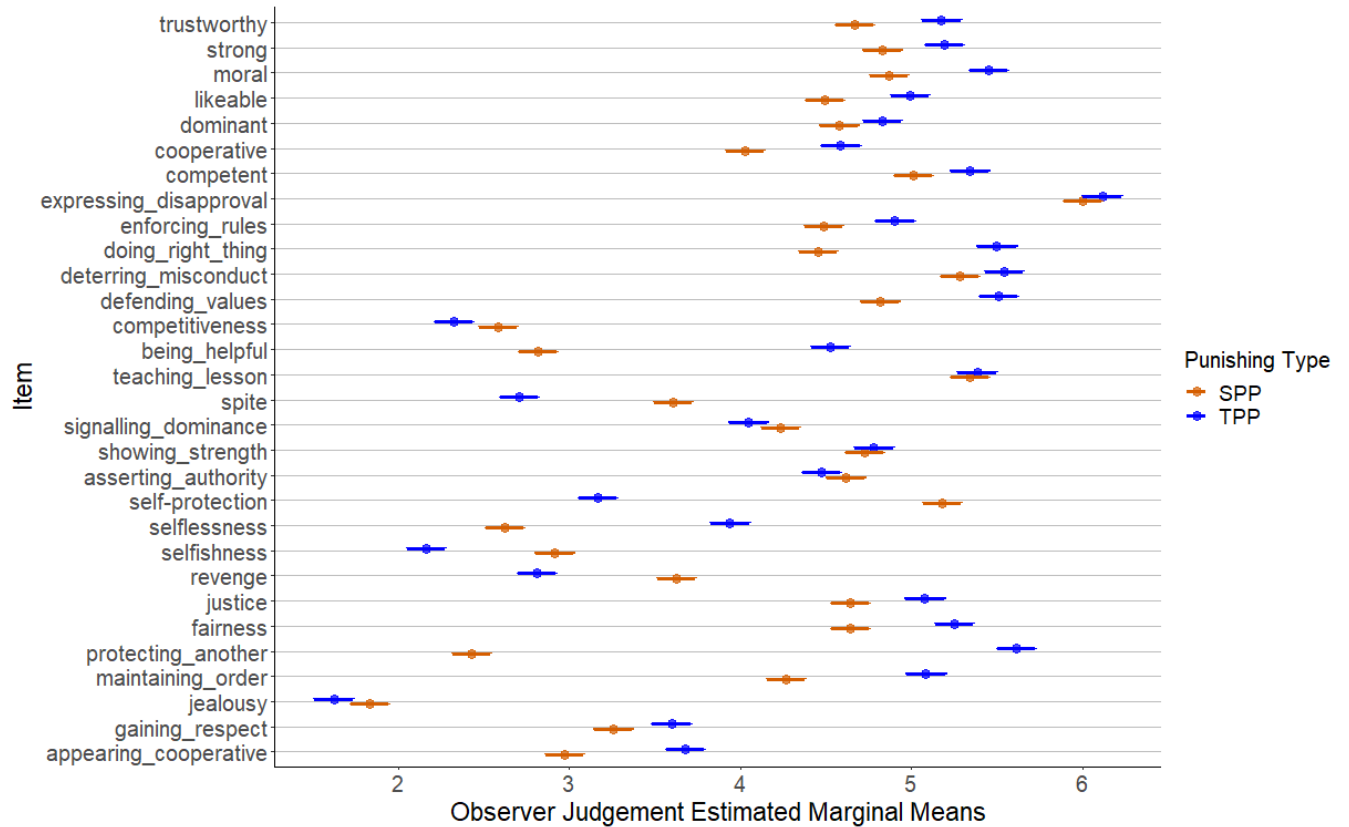
Table 5.6 illustrates alignment across observer, meta-, and self-perceptions for many motive and trait items. Across all three rating types, TPP was consistently linked more with the motives protecting others, being helpful, selflessness, doing the right thing, maintaining order, appearing cooperative, defending values, fairness, justice, and deterring misconduct, as well as the traits morality, cooperativeness, trustworthiness, likability, and competence. SPP was consistently more associated with self-protection, spite, revenge, and selfishness. Across all three perceptions types, there were no significant differences in the attribution of the motives to teach a lesson and to express disapproval between TPP and SPP.

### *Observer Perceptions*

Overall, observers' highest attributed motive for both TPP and SPP was expressing disapproval (Figure 5.10). The other top motive attributions for TPP were protecting another, deterring misconduct, defending values, doing the right thing and teaching a lesson. The other top motives for SPP were teaching a lesson, deterring misconduct, self-protection, defending values and asserting authority. The most highly rated trait for SPP was competence, and moral for TPP. The lowest rated motive for both SPP and TPP was jealousy. Observers did not think motives of selfishness, competitiveness, spite or revenge to be very applicable to TPP, nor motives of protecting others, competitiveness, selfishness and being helpful to be very applicable to SPP.

**Figure 5.10**

*Observer perceptions per items across punishing type*



Note. Dots represent estimated marginal means, while error bars indicate 95% CIs.

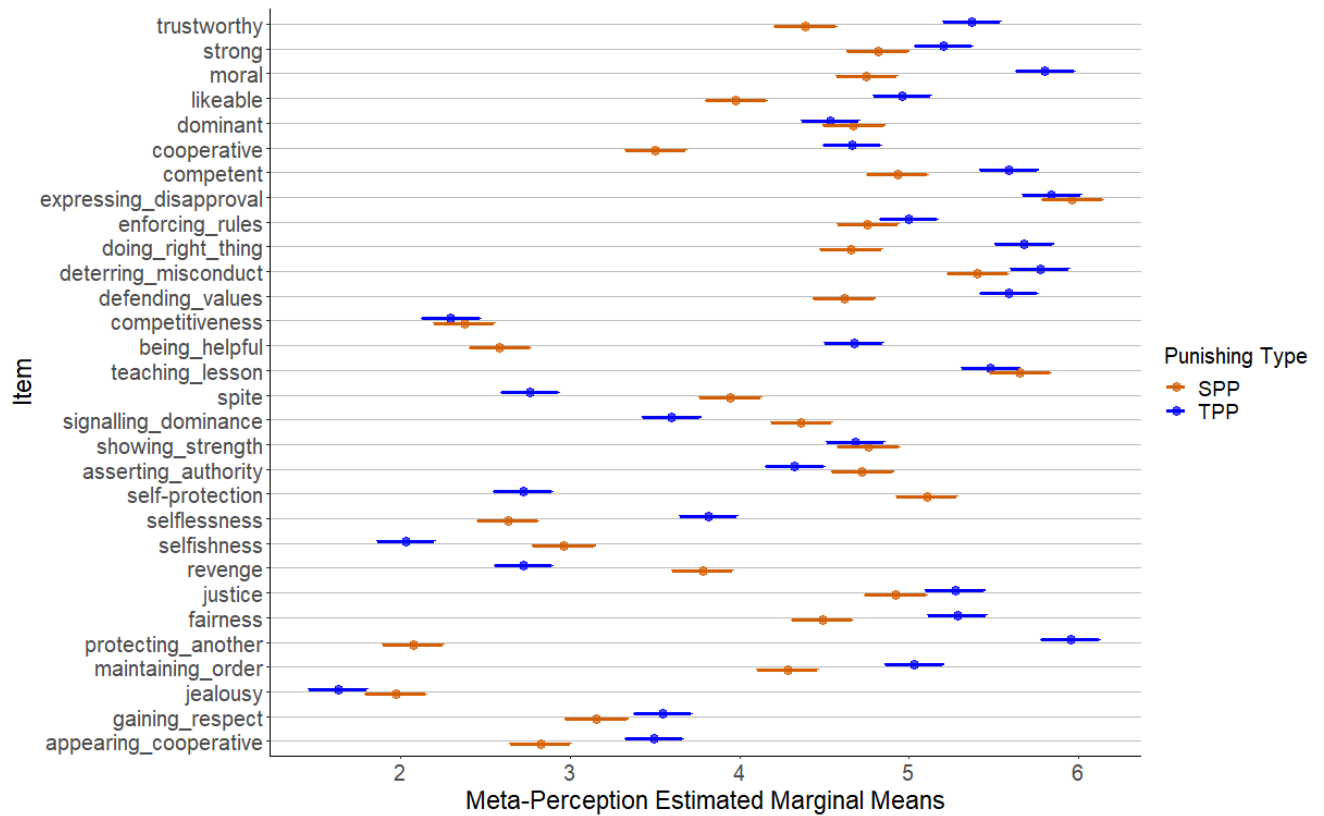
### *Meta-Perceptions*

Again, the trait which received the highest meta-perception ratings was moral for TPP contexts and competent for SPP (Figure 5.11). Actors expected their third-party punishing actions to mostly be attributed to the motives of protecting another, expressing disapproval, deterring misconduct, doing the right thing, defending values and teaching a lesson. Conversely, in SPP contexts meta-predictions were highest for the motives expressing disapproval, teaching a lesson, deterring misconduct, self-protection, showing strength and enforcing rules. Similar to observer perceptions, actors expected jealousy to be the least attributed motive in both TPP and SPP contexts. When predicting third-party punishment evaluations, actors expected observers to dismiss motives of selfishness, competitiveness, self-protection, and revenge. In second-party punishing contexts, actors expected observers to dismiss motive attributions of protecting others, competitiveness, being helpful and selflessness.



**Figure 5.11**

*Meta-perceptions per item across punishing type*



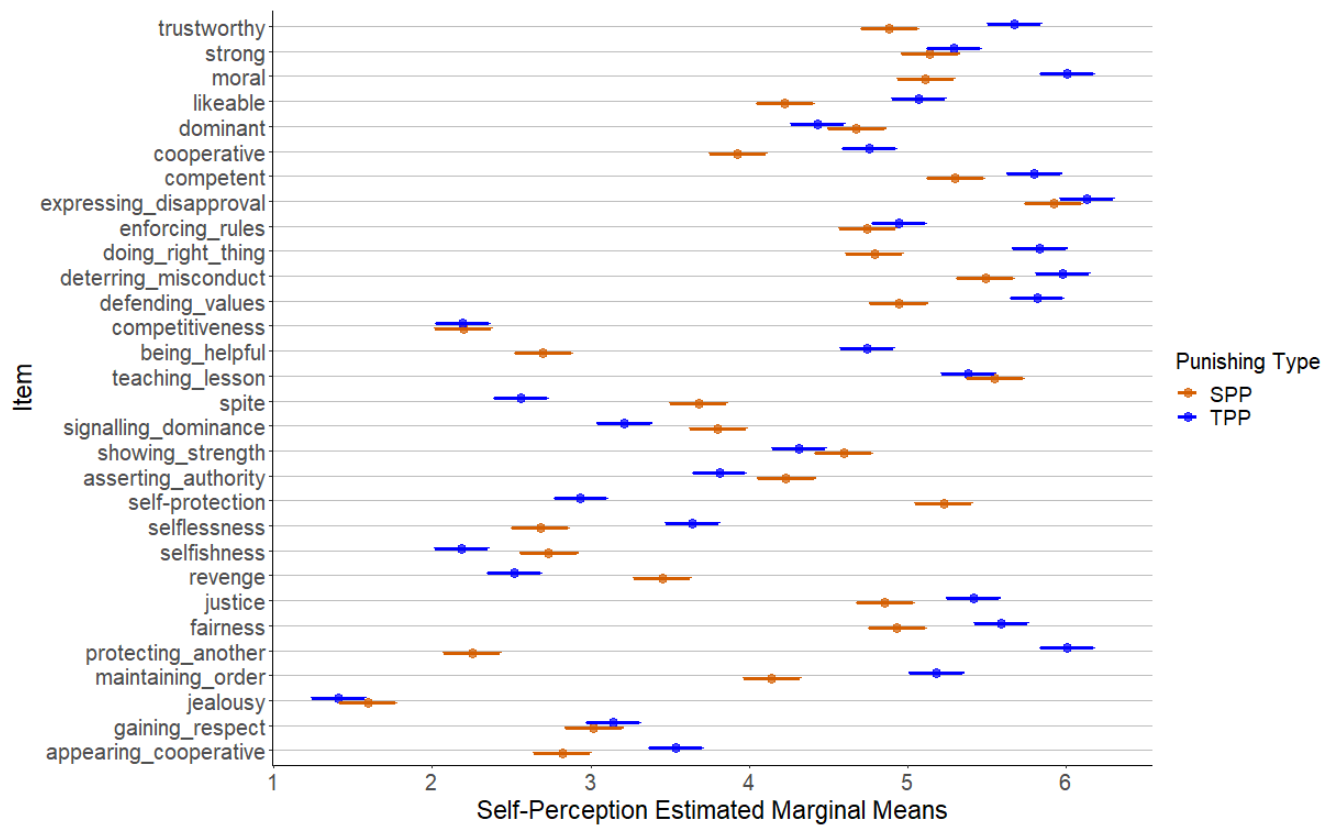
*Note.* Dots represent estimated marginal means, error bars indicate 95% CIs.

### *Self-Perceptions*

As observers expected, actors were the most motivated by expressing disapproval in both TPP and SPP contexts (Figure 5.12). Third-party punishers also tended to be motivated by a desire to protect someone else, deter future misconduct, do the right thing, defend values, justice, and teach a lesson. Second-party punishers were also highly motivated by a desire to teach a lesson and deter future misconduct, as well as self-protection, and again defending values, justice and doing the right thing. In line with observer and meta-perceptions, third-party punishers perceived themselves as especially moral, whereas second-party punishers perceived themselves as especially competent. Actors were the least motivated by jealousy in both TPP and SPP contexts, followed by competitiveness. TPP was also not particularly motivated by selfishness, revenge or spite. Conversely, SPP tended not to be motivated by protecting others, selflessness or being helpful.

**Figure 5.12**

*Self-perceptions per items across punishing type*



*Note.* Dots represent estimated marginal means, error bars indicate 95% CIs.

#### *Correlations Between Motive and Trait Items in TPP and SPP*

Following the analysis of motives distinctly associated with TPP and SPP, I examined the correlations between motive and trait items for both behaviours. The correlations provide insight into the relationships between all motives and traits in SPP and TPP contexts.

For third-party punishing events, the lowest correlations were between the items spite and protecting another ( $r = 0.00$ ), selfishness and expressing disapproval ( $r = 0.00$ ), jealousy and being helpful ( $r = 0.01$ ), jealousy and trustworthiness ( $r = -0.01$ ), selfishness and being helpful ( $r = -0.01$ ), doing the right thing and competitiveness ( $r = -0.01$ ), seeking revenge and defending values ( $r = 0.01$ ), as well as selflessness and seeking revenge ( $r = -0.01$ ). Conversely, the strongest correlations emerged predominantly between trait items. For instance, trustworthiness was highly correlated with morality ( $r = 0.78$ ), competence ( $r = 0.76$ ), likeability ( $r = 0.75$ ), and strength ( $r = 0.68$ ). Further, morality was associated with competence ( $r = 0.73$ ) and

likeability ( $r = 0.69$ ), and likeability with cooperativeness ( $r = 0.69$ ). Moreover, dominance was strongly correlated with the motivation of signalling dominance ( $r = 0.68$ ), and the motive of doing the right thing was strongly correlated with morality ( $r = 0.70$ ). Among the motive items, the strongest correlations were between signalling dominance and asserting authority ( $r = 0.68$ ), doing the right thing and asserting authority ( $r = 0.64$ ), doing the right thing and defending values ( $r = 0.63$ ), maintaining order and enforcing rules ( $r = 0.62$ ), showing strength and asserting authority ( $r = 0.62$ ), as well as selflessness and being helpful ( $r = 0.62$ ).

For second-party punishing events, the lowest correlations were between the items teaching a lesson and self-protection ( $r = 0.01$ ), restoring fairness and jealousy ( $r = -0.01$ ), deterring misconduct and competitiveness ( $r = 0.01$ ), selfishness and protecting another ( $r = 0.02$ ), seeking revenge and asserting authority ( $r = 0.02$ ), jealousy and strength ( $r = 0.02$ ), as well as jealousy and likeability ( $r = -0.02$ ). Similar to TPP events, the strongest correlations for SPP behaviours were primarily between traits. Trustworthiness was strongly correlated with morality ( $r = 0.80$ ), likeability ( $r = 0.78$ ), competence ( $r = 0.78$ ) and strength ( $r = 0.70$ ). Further, morality was associated with competence ( $r = 0.75$ ) and likeability ( $r = 0.74$ ), and strength with competence ( $r = 0.70$ ). Additionally, the motive of doing the right thing was strongly associated with morality ( $r = 0.77$ ), the motive of showing strength with the trait strength ( $r = 0.76$ ), and the motive of appearing cooperative with being cooperative ( $r = 0.76$ ). Among motive items, the strongest associations included spite and revenge seeking ( $r = 0.74$ ), dominance signalling and asserting authority ( $r = 0.74$ ), selflessness and appearing cooperative ( $r = 0.69$ ), being helpful and appearing cooperative ( $r = 0.67$ ), as well as doing the right thing and defending values ( $r = 0.67$ ).

Due to the strong similarities across the different rating types (meta-perceptions, self-perceptions, and observer perceptions), I chose to only report the correlations between motive and trait items for meta-perceptions of SPP and TPP behaviours. This decision helps streamline the presentation of results, reducing redundancy and avoiding an overload of repetitive tables, thereby making the findings more accessible. Meta-perceptions, which reflect how individuals believe they are viewed by others, may serve as an intermediary between actor self-perceptions and actual observer evaluations. Given the high consistency across the different rating types, presenting the correlations for meta-perceptions alone should adequately represent the relationships among the variables. For correlation matrices, please refer to Appendix D8.

### 5.2.3.8 Political Orientations and Punishing Sentiments

Six cumulative link mixed model were fit to examine the relationship between (i) left-right political orientation, (ii) SDO, and (iii) RWA with (i) the extent to which punishment was perceived to reflect positively on the punisher and (ii) the perceived justness of punishment. The results indicated a significant effect of left-right political orientation on reflection ( $b = -0.004$ , 95% CI  $[-0.007, -0.001]$ ,  $z = -2.62$ ,  $p = .009$ ) and justness ( $b = -0.005$ , 95% CI  $[-0.008, -0.001]$ ,  $z = -2.71$ ,  $p = .007$ ). Specifically, individuals who placed themselves more towards the right end of the political spectrum perceived punishers more negatively and punishment as less justified. Observers scoring higher on the SDO measure also believed punishing was less justified ( $b = -0.02$ , 95% CI  $[-0.02, -0.007]$ ,  $z = -3.50$ ,  $p = .0005$ ) and viewed punishers more negatively ( $b = -0.01$ , 95% CI  $[-0.02, -0.003]$ ,  $z = -2.72$ ,  $p = .006$ ). RWA did not significantly influence perceptions of punishers ( $b = 0.004$ , 95% CI  $[-0.004, 0.01]$ ,  $z = 0.91$ ,  $p = .36$ ) or perceived justness ( $b = -0.005$ , 95% CI  $[-0.01, 0.003]$ ,  $z = -1.27$ ,  $p = .21$ ).

To test whether punishers were evaluated more positively in TPP compared to SPP contexts, and whether TPP was perceived as more justified, I fit models predicting mean-centred approval (specifically whether punishing reflected more negatively or more positively on the punisher) as a function of mean-centred justness (specifically as how justified the punishing act was regarded) and punishing type, as well as their interaction, separately for observer and punisher responses. Observers believed both that punishing reflected significantly more positively on actors when they regarded the punitive action as more justified ( $b = 0.90$ , 95% CI  $[0.87, 0.93]$ ,  $t(549.21) = 63.95$ ,  $p < .0001$ ) and when the actor engaged in TPP rather than SPP ( $b = 0.32$ , 95% CI  $[0.25, 0.39]$ ,  $t(387.42) = 8.80$ ,  $p < .0001$ ). There was no interaction between justness and punishing type ( $b = 0.004$ , 95% CI  $[-0.04, 0.04]$ ,  $t(2502.34) = 0.18$ ,  $p = .86$ ). Actors, when predicting observer perceptions of approval and justness, expect to be evaluated significantly more positively when observers viewed the punishing action as more justified ( $b = 0.76$ , 95% CI  $[0.64, 0.88]$ ,  $t(282.52) = 12.25$ ,  $p < .0001$ ) and when they engaged in TPP ( $b = 0.64$ , 95% CI  $[0.38, 0.89]$ ,  $t(155.98) = 4.93$ ,  $p < .0001$ ). Again, no statistically significant interaction was found ( $b = -0.10$ , 95% CI  $[-0.30, 0.09]$ ,  $t(238.53) = -1.03$ ,  $p = .30$ ). In comparison, when actors predicted how positively they would be perceived for punishing based on their *own* perception of how justified their punishing action was and punishing type, both justness ( $b = 0.66$ , 95% CI  $[0.50, 0.82]$ ,  $t(285.31) = 7.86$ ,  $p < .0001$ ) and punishing type ( $b = 0.85$ , 95% CI  $[0.56, 1.14]$ ,  $t(150.88) = 5.72$ ,  $p < .0001$ ) again significantly predicted punisher evaluations, but the estimate for justness was comparatively lower.

## 5.2.4 Discussion

This study analysed 294 descriptions of everyday punitive actions with accompanying self- and meta-perceptions, alongside 3120 judgements on those behaviours. The primary aim was to assess whether punishers are aware of the motives observers are likely to attribute to their punishing actions (meta-accuracy), and whether observers' attributions align with punishers' reported motives and traits (observer accuracy) in the context of real-world punitive actions. By investigating whether we know how our punishing behaviours influence motive attributions and our reputation, I aimed to further our understanding on the dynamics of reputation management and interpersonal communication within punitive contexts.

Using vignettes, the first study of this Chapter found that observers associated TPP more with prosocial motivations and SPP more with self-protection and formidability concerns, as well as perceiving TPP more positively than equivalent SPP actions. One aim of the current study was to confirm whether these findings could be replicated with a more complex method, in which experimental control was reduced in favour of enhanced ecological validity. Indeed, observers again attributed prosocial motives (e.g., protecting another, being helpful, selflessness, and doing the right thing) as well as reputation benefits (e.g. traits moral, cooperative, trustworthy, and likeable) to TPP over SPP. Unexpectedly, reputation benefits were not limited to prosocial aspects, as third-party punishers were also perceived as stronger, more competent, and more dominant. In line with predictions and findings from the vignette study, observers again attributed selfish and formidability concerns to second- over third-party punishers (e.g. self-protection, spite, revenge, selfishness, and competitiveness). As found in the first study of this Chapter, observers did not significantly differ in attributing the motives of expressing disapproval and teaching a lesson to SPP or TPP. On average, the highest attributed motive for both TPP and SPP was expressing disapproval. The motives teaching a lesson, defending values, and deterring misconduct were also, on average, highly attributed to punishers in both contexts. This study therefore recorded similarly distinct punishing motive attributions as observed in the vignette study. In addition, I showed that punishers expectations match this divide of motive and trait attribution: they anticipated that prosocial motives and traits would be attributed to their TPP over their SPP actions, and self-focused and formidability related motives would be seen as more applicable to their SPP actions. Actors' self-reported motives and traits followed the same patterns as well. On average, the highest expected and reported motives by actors also matched observer perceptions, as the motives expressing disapproval, teaching a lesson, and deterring misconduct were highly rated for both SPP and TPP. These

motives relate to the overarching goal of reinforcing societal norms and sanctioning deviant behaviour, reflecting individuals' efforts to regulate behaviour in accordance with the prevailing norms and standards of their society. Jealousy and competitiveness were consistently dismissed as relevant motives for both TPP and SPP, across all ratings. The most highly rated trait was always moral for TPP and competent for SPP.

Before punishing, individuals often consider how their punitive action could influence observers' perception of them and their motives (Radkani et al., 2022). This aligns with expectations that punishment serves as a communicative action (Cushman et al., 2019; Funk et al., 2014; Ho et al., 2019, 2021; Molnar et al., 2020; Sarin et al., 2021). For punishment to be a successful communicative action both actors and observers must accurately understand the relationship between the punitive actions and the motives behind them, as well as the inferences drawn by observers (Sarin et al., 2021). Consequently, I anticipated participants would provide accurate observer and meta-perceptions for both second- and third-party punishing behaviours.

To directly test for this I adopted a componential approach to modelling judgment accuracy (Biesanz, 2010; Funder, 1995; Kenny, 2004; Lees et al., 2022; West & Kenny, 2011). This approach allows for the disentanglement of processes contributing to social judgment, controls for confounds present in the literature, and enables the modelling of within-person accuracy to examine moderating traits. All components of judgment are conceptualized as linear, reflecting knowledge of the rank-order of others' perceptions. Additionally, I examined mean point-estimate accuracy of single perceptions to reveal observer and meta-biases.

In line with predictions, punishers displayed meta-perceptive accuracy across all models. In both second- and third-party punishing contexts, actors exhibited knowledge of how the average actor was perceived (normative accuracy; Wood & Furr, 2016), how they were uniquely perceived (distinctive accuracy; Biesanz, 2010; Furr, 2008), and how they might be misperceived by observers (insight; Carlson, 2016). At the same time, actors still overestimated how accurately observers could perceive them (transparency bias; Gilovich et al., 1998). This was further shown through the high correlations between actors' meta- and self-perceptions, which suggest that punishers did not expect observers' perceptions to deviate much from their own.

Similarly, observers were able to accurately assess second- and third-party punishers' actual motives and trait self-perceptions, exhibiting baseline, distinctive and normative accuracy. They also demonstrated insight, by accurately assessing actors' actual motives and traits beyond how actors expected to be perceived. At the same time, the observer insight model also found

evidence for opacity, which means that observers perceived some items inaccurately and in a way that actors anticipated. This suggests that individuals are able to successfully perspective-take with punishers and punishers can anticipate how they would be perceived by observers.

As expected, punishing type moderated meta-perceptive accuracy. Punishers were significantly more accurate in predicting how they would be perceived for their second- compared to their third-party punishing actions (although normative accuracy was higher for TPP). It is possible, that due to SPP being more prevalent than TPP (Kriss et al., 2016), individuals have more exposure to and therefore experience with SPP. Both from observation and personal experience, people may therefore be more accurate at predicting how their punishing action would be perceived by others. Additionally, SPP motives might be regarded as and actually be more straightforward (e.g. self-protection) whereas seemingly prosocial TPP actions could involve additional ulterior motives. As expected, observers also demonstrated significantly higher levels of accuracy when evaluating second- compared to third-party punishers.

For observers, extroversion, agreeableness, conscientiousness, neuroticism, openness, empathic concern, and perspective taking did not moderate distinctive observer accuracy. For punishers, the personality traits agreeableness, conscientiousness, neuroticism, and openness also did not moderate meta-accuracy. While extroversion did not moderate meta-accuracy in TPP contexts, more extroverted second-party punishers displayed significantly lower levels of meta-accuracy. This is contrary to predictions, which were based on the recent finding that individuals with strong meta-perception skills in first impressions tend to be extroverted (Hater et al., 2023). A possible explanation for the current findings could be, that extroverts typically possess higher levels of self-confidence (Francis, 1997; Kawash, 1982; Schmitt & Allik, 2005; Swickert et al., 2004). This confidence may lead them to believe that their actions are justified and acceptable, regardless of how others might perceive them, resulting in lower meta-accuracy in predicting others' evaluations. Additionally, extroverts tend to engage in more risky behaviours compared to introverts (Desmoulins-Lebeault et al. 2018; Eysenck, 1976; Tassone & Foster, 2021; Zuckerman & Kuhlman, 2000). This propensity for risk-taking might lead extroverts to underestimate how their second-party punishing actions will be perceived by others, as they may be less cautious or reflective about potential consequences. Conversely, introverts may be more accurate than extroverts because they engage in less risky, more common punishing behaviours, where it is easier to predict how others will react based on experience, or only punish when they are fairly certain that they can anticipate what response

they will receive. Alternatively, introverts may exhibit higher meta-accuracy because they are more comfortable with online interactions and self-disclosure (Amichai-Hamburger et al. 2002), which may have led to richer behaviour descriptions from which more precise predictions could be made. Furthermore, a recent study found that extroverts viewed themselves as more similar to their peer group than introverts did (Vaughan-Johnston et al., 2021), which could explain why extroverts would struggle to display distinctive accuracy.

Further, in SPP contexts, empathic concern negatively influenced distinctive meta-accuracy. Empathic concern involves identifying with and understanding the emotions of others. However, this may lead individuals to adopt the perspective of the target of their punishment rather than an objective observer. As a result, they may struggle to distinctly assess how their actions will be perceived by others who are not directly involved in the situation. Alternatively, high levels of empathic concern may lead individuals to prioritize emotional responses over cognitive evaluations when assessing how their actions will be perceived. This emotional bias could interfere with their ability to accurately predict how others will view their behaviour in particular. Normative accuracy, on the other hand, was positively influenced by both empathic concern and perspective taking in SPP and TPP contexts. Due to reports of impaired meta-accuracy in autism (Sasson et al., 2018), researchers have suggest that a positive association between social competencies and meta-accuracy is probable (Tsankova & Tair, 2021). Social skills may therefore improve actors' knowledge of how the average punisher is perceived.

Furthermore, I expected punishers to believe others would judge them less favourably than they actually do, indicating a systematic negative meta-bias. Contrary to predictions, third-party punishers exhibited a positive meta-bias overall. More specifically, they overestimated how much their third-party punishing action would positively influence their prosocial reputation, and underestimated how much TPP would influence their formidability reputation. Actors for instances expected to be viewed as more moral and trustworthy than they actually were perceived, and overestimated prosocial motive attributions such as protecting others, doing the right thing, bringing justice or being helpful. Simultaneously, they underestimated the attribution of more formidability and self-focused motives such as signalling dominance, self-protection, asserting authority, or revenge. However, the positivity bias here should be interpreted in context: for the items that third-party punishers over-estimated, they knew that observer perceptions would be lower than their self-perceptions – actual perceptions were just even lower than actors anticipated. Mirroring this, for under-estimated items actors generally expected observer attributions to be higher than their self-perceptions. The positivity bias



should therefore not be viewed as a mere positivity bias in self-evaluation, as third-party punishers did have knowledge of the directional difference in observer- versus self-perceptions.

In contrast, there was no systematic bias in second-party punishing situations. Overall, actors both *underestimated* the positive effects of second-party punishment on their prosocial reputation (e.g., expecting to be viewed as less cooperative, likeable, trustworthy or less motivated by protecting others and being helpful than they really were) and *overestimated* how much observers would attribute their punishing action to principles of group living and fairness (e.g., by expecting their actions to be perceived as motivated by justice, enforcing rules, teaching a lesson and doing the right thing).

An additional unregistered analysis tested for observer bias. Overall, observers exhibited a strong negativity bias in both TPP and SPP contexts, believing that punishers were more motivated by formidability or self-centred concerns than they actually were (e.g., overestimating how much punishment was motivated by a desire to signal dominance, assert authority, show strength, gain respect, or competitiveness) and thinking punishers were less prosocial and prosocially motivated than actors indicated (e.g. rating punishers as less moral and trustworthy than actors perceived themselves, and underestimating motives such as fairness, justice, doing the right thing, deterring future misconduct, or being helpful). Altogether, observers were more suspicious of punishers and underestimated their other-regarding motives whilst over-attributing more self-centred motives. These findings are in line with prior research which highlight that punishing often improves the punisher's situation in comparison to that of the violator, leading observers to doubt the moral legitimacy of punishment (Fehr & Rockenback, 2003; Raihani & Bshary, 2015b; Xiao, 2013).

As punishers are generally viewed more positively when their actions stem from cooperative rather than selfish motives (Barclay, 2006; Jordan et al., 2016; Kiyonari & Barclay, 2008; Raihani & Bshary, 2015b,c), and given that TPP is more likely to involve cooperative motivations than SPP (Carpenter, 2003; Espin et al., 2012; Peysakhov et al., 2014; Yamagishi et al., 2012), I anticipated that third-party punishers would receive more favourable evaluations. Furthermore, the perception of punishment is influenced by the perceived deservedness of punishment (Salcedo & Jimenez-Leal, 2023). In line with this, punishers were indeed evaluated more positively when they engaged in TPP compared to SPP, and when the punishment was perceived as more justified. Actors accurately predicted this pattern. Moreover, actors anticipated that their TPP actions would be perceived as more justified than their SPP behaviours, which mirrored observers actual evaluations.

Lastly, political orientations somewhat influenced punishing sentiments. More conservative individuals - those towards the right end of the political spectrum and those scoring higher on SDO - perceived punishers more negatively and punishment as less justified. RWA did not significantly influence perceptions of punishers or justness. This finding stands in contrast to prior research which reports that conservatives punish, or desire to punish, perpetrators more harshly than individuals with other political orientations do (Carroll et al., 1987; Clark & Wink, 2012; Hoffmann et al., 2018; Hogan et al., 2005; Johnson, 2009; King & Maruna, 2009; Miller, 1973; Robinett & Malhotra, 2019; Tetlock et al., 2007; Tyler & Boeckmann, 1997). An explanation for these findings has been that conservatives are more concerned with personal agency than external circumstances (Schlenker et al., 2012; Weiner et al., 2011) whilst liberals often believe in economic and external causes of crime and therefore prefer rehabilitation over punishment when deliberating about sentencing goals (Carroll et al., 1987). Perhaps the influence of political orientation on punishing sentiments is context-dependent, differing between official, more serious sentencing decisions and punitive actions taken in everyday life. In fact, punishment does not serve just one function (Raihani & Bshary, 2019), but is a flexible tool that is used for different purposes (Claessens et al., 2024 Preprint). A recent study reported that political ideology influences punishing strategies, for instance finding that conservatives were more focused on avoiding disadvantageous inequity and less inclined to follow egalitarian strategies (Claessens et al., 2024 Preprint).

It is essential to recognise that findings may not necessarily apply to the individuals being punished as perpetrators or the victims in TPP scenarios. Observers served as neutral third parties, and actors were prompted to provide meta-perceptions of "the average person," not the violator, victim, or any individuals directly involved in the narratives. Consequently, these results should not be construed as evidence that punishers possess insight into how violators or victims perceive their actions or that 'observers' would maintain rational accuracy when personally involved (i.e., violators and victims may not be as accurate at guessing the motives underlying a punisher's actions).

Additionally, the contexts in which SPP and TPP occur may differ in ways that go beyond whether the punisher was the victim. Such contextual variations can affect how punitive behaviours are reported and perceived. For instance, SPP incidents may involve different relational dynamics or thresholds for engagement compared to TPP incidents. SPP tends to be more common than TPP, as individuals are generally more inclined to seek out opportunities for personal retribution, while third parties engage in punishment more hesitantly (Kriss et al.,

2016). This comparatively greater frequency of SPP may give participants a wider array of instances to report, potentially allowing them to choose incidents that cast them in a more favourable light or that seem more interesting. Additionally, people use various punitive strategies (Balafoutas et al., 2014, 2016; Molho et al., 2020), some of which may differ in likelihood between TPP and SPP. Indeed, research on direct (e.g., confrontation) and indirect punishment (e.g., gossip, social exclusion) in everyday life shows that direct confrontation is more prevalent in SPP contexts (Molho et al., 2020).

Furthermore, the personal involvement in SPP versus the distance in TPP could have impacted participants' reporting styles, with SPP potentially being reported with different emotional content than TPP. Participants may recall instances of SPP with heightened emotional intensity due to a focus on personal grievances, leading to a possible overreporting of aggressive or severe punitive behaviours. In contrast, TPP may be focused on a sense of moral obligation or social norm enforcement, which may lack the intensity of SPP. Additionally, participants often justified their punishing actions, which might have similarly differed between SPP and TPP reports. For instance, SPP actions could carry implicit biases for self-defence, with more aims to frame their punishment as deserved or justified. By contrast, TPP may be presented as more altruistic or noble, with participants perhaps more likely to frame these actions around defending fairness or others. Observers might then rate SPP actions differently from TPP not only due to the punishment act but also because of how participants frame their involvement. See Appendix D9 for a sample of participant responses, providing insight into the range of second- and third-party punishing behaviours reported.

In conclusion, this study provides insights into the dynamics of reputation management and interpersonal communication within real-world punitive contexts. Through an analysis of everyday punitive actions I showed that observers tend to associate TPP with prosocial motivations and traits, compared to SPP which was attributed more self-focused motives. Actors accurately predicted this pattern, anticipating that their TPP actions would be viewed more favourably than their SPP behaviours. Reputation benefits were not only stronger in TPP compared to SPP contexts in the domain of prosocial reputation, but also saw more formidable traits attributed to third-party compared to second-party punishers. Punishers and observers consistently demonstrated meta-perceptive and judgment accuracy in both TPP and SPP contexts, with punishers correctly predicting how they would be evaluated for their punishing behaviours, and observers accurately attributing motives and traits to punishers. At the same

time, observers and punishers still exhibited biases, and there were higher levels of accuracy observed for SPP compared to TPP.

# Chapter 6

## **Deliberating Cost and Impact: Trustworthiness Signals in Punishment and Helping**

### **STATEMENT OF PUBLICATION**

The results presented in this chapter were published in *Communications Psychology*.

• Engeler, N. C., & Raihani, N. J. (2024). Acting Without Considering Personal Costs Signals Trustworthiness in Helpers but Not Punishers. *Communications Psychology*, 2(1), 47. <https://doi.org/10.1038/s44271-024-00092-7>.

## 6.1 Introduction

As outlined in previous chapters, prosocial behaviours, such as helping and cooperating, can benefit others but often come at a personal cost to the actor (Fehr & Fischbacher, 2003; Penner et al., 2005; West et al., 2007). Punishment, which involves an actor paying a cost to impose a cost on a social partner (Clutton-Brock & Parker, 1995; Raihani & Bshary, 2019), can encourage and maintain prosocial behaviours by deterring selfish actions (Balafoutas et al., 2014; Balliet et al., 2011; Boyd & Richerson, 1992; Boyd et al., 2003; Charness et al., 2008; Feinberg et al., 2014; Henrich & Boyd, 2001; Matthew & Boyd, 2001). Although punishing anti-social behaviours can increase group-level cooperation, it also imposes a cost on the punisher by requiring effort and time, and puts the punisher at risk of retaliation (Balafoutas et al., 2014; Dreber et al., 2008; Nikiforakis, 2008). To understand why people invest in punitive acts, we must explain how punishment might ultimately lead to downstream benefits for the punisher.

This question is particularly pertinent when it comes to third-party punishment (TPP), where a punisher intervenes to punish a cheat even though they were not personally harmed by the cheat's behaviour and may not interact with the target of punishment again in the future. Third-party punishment can still provide reputation benefits to the punisher (Barclay, 2006; dos Santos & Wedekind, 2015; dos Santos et al., 2013; Kurzban et al., 2007; Raihani & Bshary, 2015b), either by signalling their formidability (which may deter their current social partners or bystanders from transgressing in the future; dos Santos et al., 2013; Hilbe & Traulsen, 2012) or by signalling their cooperative intent (which may result in others being more likely to cooperate with them (Raihani & Bshary, 2015c) or choose them as partners for cooperative interactions (Barclay, 2006; Batistoni et al., 2022; Dhaliwal et al., 2020; Jordan et al., 2016; Nelissen et al., 2008). Third-party punishment can therefore act as a signal that communicates an otherwise unobservable intent to act prosocially (Horita, 2010; Jordan & Rand, 2017; Jordan et al., 2016; Nelissen, 2008; Przepiorka & Liebe, 2016; Raihani & Bshary, 2015b, 2015c). Accordingly, in some settings, individuals invest more in third-party punishment when they are observed (Batistoni et al., 2022; Kurzban et al., 2007) and are evaluated in a preferential manner by others for doing so (Jordan et al., 2016).

Nevertheless, punishment is, by definition, a harmful act, which complicates inferences about the punisher's intentions (Batistoni et al., 2022; Raihani & Bshary, 2015b, 2019). Punishment could stem from antisocial, competitive, or spiteful motivations rather than from a desire to cooperate, promote fairness, or uphold social norms (Raihani & Bshary, 2019). Indeed,

compared to those who compensate victims, third-party punishers typically have higher scores for antisocial personality traits such as Machiavellianism, narcissism, and psychopathy (Dhaliwal et al., 2021). Third-party punishment is therefore a more ambiguous signal of trustworthiness and cooperative intent than helping or compensating a victim (Batistoni et al., 2022; Dhaliwal et al., 2021; Heffner & FeldmanHall, 2019; Jordan et al., 2016; Raihani & Bshary, 2015b); and is most likely to signal cooperative intent in scenarios where the punisher cannot compensate the victim, and where self-serving motives are less likely (Raihani & Bshary, 2015b) (e.g. when the punisher does not increase their own payoffs relative to those of the target when they punish; Raihani & Bshary, 2015b, 2019).

The potential for helpful acts to signal cooperative intent also depends, to some extent, on context (Raihani & Power, 2021). One such context concerns whether the helpful act was calculated or not: uncalculated help is a stronger signal of the helper's cooperative disposition than calculated help. One recent study operationalised uncalculated help by measuring whether individuals chose to learn the personal cost to themselves before helping, and by recording how long it took individuals to make their helping decision once the cost of helping was revealed (Jordan et al., 2016). Helpers who check the cost are ostensibly weighing up the costs and benefits of their actions, suggesting that cooperating is a more strategic and calculated move. Response time is also informative about a person's underlying commitment to cooperation, as slower decisions indicate greater decision conflict (Nishi et al., 2016). Observers use information about another's decision time to infer levels of conflict experienced and to make predictions about whether a person is making a calculated or uncalculated decision to cooperate (Cricher et al., 2013; Evans et al., 2017; Jordan et al., 2016; Van de Calseyde, 2014), and whether to trust them (Jordan et al., 2016). This previous work found that uncalculated cooperation was a more reliable signal of trustworthiness than calculated cooperation (Jordan et al., 2016). Moreover, people were apparently aware of the signalling value of uncalculated cooperation and were less likely to check the costs of cooperating and made cooperative decisions more quickly when observed.

Although both third-party punishment and helping are prosocial acts, decision conflict for these behaviours might not be interpreted in the same way, and it is therefore unclear whether findings on uncalculated help (Jordan et al., 2016) would be expected to translate directly to the punishment setting. I address this issue here. Decision conflict over whether to help another is likely to stem from self-interested considerations of whether to pay a personal cost. Decision conflict over punishment, by contrast, could also stem from concerns about inflicting harm on

the target. This perspective yields nuanced predictions about the two different measures of decision conflict above. Checking the personal cost of administering punishment is likely to indicate a self-interested concern about personal costs. As with helping, such calculated decisions may be perceived negatively by observers. However, because punishment involves imposing costs on another individual, taking longer to decide whether to punish another person might not be viewed negatively and could even be viewed positively. Perhaps carefully thinking about and balancing both the prosocial aspect as well as the negative consequences for the punished is the ‘right’ thing to do when deciding whether to engage in third-party punishment.

To better define the conditions under which punishers are viewed positively and to differentiate between punishment and helping as signals of trustworthiness, I conducted two studies. Study 1 aimed to replicate Jordan et al.'s (2016) research on uncalculated helping and extend it to punishment, asking how deliberation over personal costs affects trustworthiness perceptions in both cases. Study 2 extended this by asking how deliberation over the impacts on targets affects perceptions of trustworthiness for both punishment and helping. With this approach I hoped to understand whether and why punishers are evaluated differently to helpers, and to show how deliberation differentially signals trustworthiness in decisions to punish or help others. Overall, I expected to show that deliberating about the *personal cost* of one's actions signals untrustworthiness for both punishing and helping, whereas deliberating about the *impact* of those actions on others constitutes the differentiating factor between helping and punishing behaviours. Specifically, I expected that those who deliberate about the impact of punishment are viewed relatively positively, whereas those who deliberate about the impact of helping are viewed relatively negatively. See Appendix E1 to view a description of all hypotheses, as preregistered with *Communications Psychology*.

Across two studies, comprising five experiments (Table 6.1), I investigated whether and when uncalculated punishment and help are used as signals of trustworthiness. As in Jordan et al. (2016), all experiments had two stages: a first stage, where Player A could pay a cost to help a victim / punish a cheat; and a second stage, where Player B decided whether and how much to trust Player A. Any money entrusted by Player B was tripled by the experimenter and Player A then decided how much to return to Player B, yielding a measure of trustworthiness. I included two conditions: Player B was either able to make their trusting decision based on (i) Player A's decision and decision process in the first stage (via cost-checking / decision time), or (ii) solely on Player A's help / punishment decision, with the decision *process* remaining concealed.



**Table 6.1**

*Nomenclature for studies investigating whether helping and punishment decisions signal trustworthiness.*

| <b>Study Identification</b>                   |                                  |                                  |                                               |                         |
|-----------------------------------------------|----------------------------------|----------------------------------|-----------------------------------------------|-------------------------|
| <b>Study 1<br/>Personal Cost Deliberation</b> |                                  |                                  | <b>Study 2<br/>Target Impact Deliberation</b> |                         |
| <b>1.1<br/>Help</b>                           | <b>1.2<br/>Punish</b>            |                                  | <b>2.1<br/>Help</b>                           | <b>2.2<br/>Punish</b>   |
| cost checking<br>(E1)                         | 1.2a<br>cost<br>checking<br>(E2) | 1.2b<br>decision<br>time<br>(E3) | impact checking<br>(E4)                       | impact checking<br>(E5) |

*Note.* I recruited 1,306 Player A - Player B pairs for each of the five experiments above (i.e., each experiment contains 1,306 Player As and 1,306 Player Bs). In each experiment, half of the players were assigned to the observable decision process condition, while the other half was assigned to the hidden decision process condition.

Four of the five experiments in this chapter operationalised deliberation through cost- or impact-checking behaviours. Due to financial constraints, I included only one decision time study, specifically focussing on personal costs of punishment. This context holds particular significance, as I expected to observe significant differences between decision time and cost checking. Longer decision times may be attributed to concern for the target, in addition to self-interested considerations. Because decision conflict over punishment could also stem from concerns about inflicting harm on the target, this experiment forms a bridge between Studies 1 (personal cost deliberation) and 2 (target impact deliberation).

Study 1 investigated whether and how personal cost deliberation signals trustworthiness. Calculated behaviours were operationalised as decisions to check the personal cost of helping / punishing (1.1 & 1.2a), or long decision times after the cost of punishing is revealed (1.2b). Study 1 therefore comprises three separate experiments (Table 6.1), investigating how deliberating over decisions to help (1.1) or to punish (1.2a & 1.2b) signal trustworthiness.

Study 2 was similar to Study 1 but investigated whether deliberation over impacts on targets of help / punishment signals trustworthiness. In Study 2, the personal cost of help / punishment

was therefore known. Calculated behaviours were instead defined as decisions to check the impact of help / punishment ('impact checking', 2.1 & 2.2).

When deliberation was operationalised as checking the personal cost of helping / punishing (Experiments 1.1 and 1.2), I predicted that uncalculated decisions signal trustworthiness: individuals who do not check the cost of helping / punishing would be entrusted with more money than those who do check the cost. I also expected participants to be sensitive to these potential reputation benefits and to be less likely to check the cost of helping / punishing when they were observed than when their decision process was hidden from Player B. Further, I expected that individuals who do not check the cost of helping / punishing would be more trustworthy than those who do check the cost. Lastly, I anticipated uncalculated help to be a stronger signal of trustworthiness than uncalculated punishment.

When personal cost deliberation was operationalised as decision speed (Experiment 1.2b), predictions were more nuanced. Here, the cost of punishing was revealed to Player A right before they made their decision. As the sole new information provided to participants prior to measuring their decision-making time related to the personal cost of punishing, I predicted that here too uncalculated punishment would be used as a signal of trustworthiness. Specifically, I expected that participants would exhibit faster decision-making when their decision time was revealed to others, in contrast to when the decision process was concealed and unable to confer any reputation-related advantages, as was observed with helping in Jordan et al. (2016).

Nevertheless, I also envisaged some differences in how quick decisions to punish might be perceived by others, which would be driven by the different motives attributed to punishers and helpers. Jordan et al. (2016) found that fast decisions to help others were perceived positively, likely because fast decisions were associated with less decision conflict. However, because punishment inflicts harm on the target, fast decisions to punish could be evaluated differently (Raihani & Bshary, 2015b, 2019). One possibility is that observers interpret a fast decision to punish as the punisher being moral and interested in restoring fairness regardless of the cost to the self. Here, uncalculated punishment would be approved of, and observers would infer that fast punishers were more trustworthy than slow punishers. Another possibility, however, is that observers may approve of more considered decisions to punish others, if they infer that decision conflict stems from concern about the harm caused to the target. Thus, slow punishers may not be evaluated as negatively and, consequently, I expected decision speed to be a weaker signal of trustworthiness than cost-checking decisions for punishment.

Despite this ambiguity, I still expected fast punishers to be trusted more than slow punishers in the context of personal cost deliberation. This is because observers were informed that the only new information Player A received right before making their punishing decision was the cost of punishment to themselves. Observers should infer, therefore, that deliberation stems only from the consideration of this personal cost and not the impact to the target. In addition, to disambiguate personal costs from harm aversion, I set the minimum potential cost of helping or punishing to be £0.00 in Study 1. Therefore, I anticipated that observers would send more of their endowment to third-party punishers who made their decision quickly, using uncalculated third-party punishment as a signal of trustworthiness. Similarly, I believed that being slower in the decision to punish would reflect the punisher's conflict about whether paying the cost would be beneficial to themselves, rather than an additional consideration of whether harming the violator is the "right thing to do". Thus, I expected uncalculated (fast) punishers to return more money than calculated (slow) punishers.

Study 2 was designed to address some of the open questions raised by Study 1 – specifically, whether deliberating over the impact on targets is perceived more positively for punishment than for helping. For both helping and punishment, I expected participants to be sensitive to the reputational consequences of their behaviour, albeit in different ways. If they want to be evaluated positively by an observer, helpers should be less likely to check how much helping will impact targets, whereas punishers should be more likely to check how much their actions will harm another. In other words, unlike Study 1, calculated punishment now served as a signal of trustworthiness: punishers who deliberate about the impact to the target should be entrusted with more money by observers and should be more trustworthy, compared to punishers who do not deliberate in this way. For helpers, as in Study 1, I expected uncalculated decisions to signal trustworthiness.

I also had several secondary predictions pertaining to decisions *not* to help or punish that can further clarify when and why deliberating over social actions carries reputation consequences. Specifically, I was interested in whether non-punishers are evaluated differently to non-helpers – and to what extent deliberation moderates these perceptions.

To understand how non-helpers and non-punishers are perceived, we must consider the potential motives driving decisions not to help or punish others. One primary reason individuals may refrain from helping or punishing in this task is due to personal costs. Additionally, non-helpers may not be especially motivated to help others because they are antisocial or inequity averse (i.e., they do not want someone else to receive more than them). For punishment,

individuals may also refrain because they are averse to harming others. To disambiguate personal costs from harm aversion, I set the minimum cost of helping or punishing to be £0.00 in Study 1 (the minimum impact of helping or punishing was set to £0.01, so that some impact of investing in help or punishment was guaranteed). This feature allowed me to make nuanced predictions about how non-helpers and non-punishers would be evaluated.

A decision not to help is likely to stem primarily from self-interest and, consequently, unhelpful individuals are generally not trusted by others (Batistoni et al., 2022; Jordan et al., 2016). Those who decide not to help without checking the cost to themselves (Experiment 1.1) or the impact to the target (Experiment 2.1) might be evaluated especially negatively as it indicates an unwillingness to help, even if helping might impose no personal cost (Study 1) and regardless of the potential benefit to the target (Study 2). Conversely, deliberation indicates that the individual at least considered helping before deciding not to. Thus, in general, uncalculated decisions not to help should be evaluated negatively by observers in both studies, and uncalculating non-helpers should be less trustworthy than calculating non-helpers.

The same is not true for punishment: refraining from punishing others could stem from self-interest or from harm aversion. The possibility for non-punishment to stem from harm-aversion may help to explain why non-punishers can sometimes be trusted as much as punishers (Batistoni et al., 2022). In Study 1, I expected that non-punishers who do not consider the personal cost might be perceived as harm averse – individuals who would not punish even if it were free to do so. Here, non-punishers who do not check the cost of punishing might be perceived as (and should actually be) relatively trustworthy. Conversely, non-punishers that check the personal cost before refraining from punishment should be seen as less trustworthy, because the inference is that these decisions were driven by self-interest (i.e. the personal cost being too large) rather than harm aversion.

In Study 2, uncalculated non-punishers (those who did not check the impact of punishment on targets) might either be completely harm averse, or might be unwilling to pay the personal cost associated with punishing. Calculated non-punishers (those who *did* check the impact of punishing) by contrast, are those who might be willing to incur the personal cost of punishing but who wanted to know what impact this would have on the target before doing so. For uncalculated non-punishment to be perceived as more trustworthy than calculated non-punishment, participants would need to believe the target deserved no punishment. However, this would also imply tacit acceptance of the behaviour exhibited by Player 2 (returning nothing after their partner entrusted them with their entire endowment). Since both Players A and

Players B knew they would subsequently be playing a Trust Game together, attitudes towards the target and what is considered acceptable behaviour in a Trust Game are relevant. As such, I expected calculated non-punishers to be perceived as and actually be more trustworthy than uncalculated non-punishers.

Given that the motives for actions are somewhat more transparent than those for non-actions, and that I expected incurring a cost to punish the cheat or help the cheated to be seen as more prosocial than doing nothing, I anticipated that deliberation (both of personal cost and target impact) would have a more substantial impact on trust and trustworthiness when individuals chose to punish or help than when they did not. Finally, I again expected non-punishers' decision speed to have a weaker effect on trust and trustworthiness than cost-checking decisions.

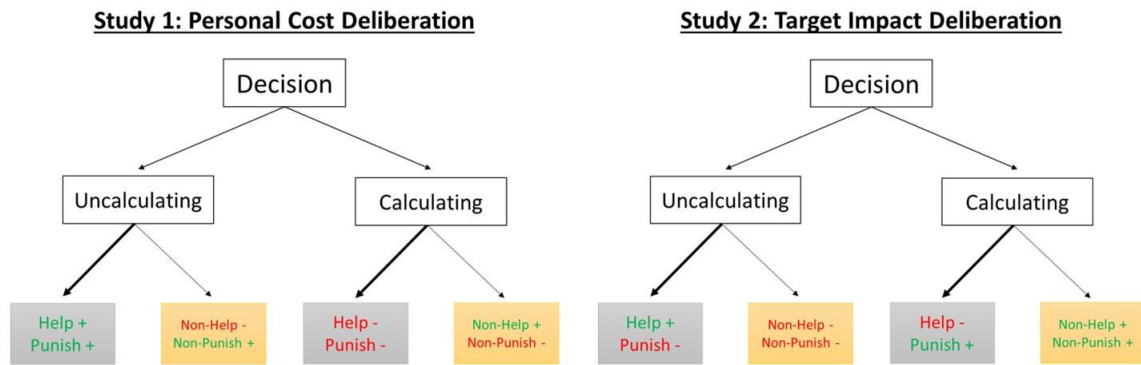
I must note, however, that predictions regarding the trustworthiness of punishers/helpers and non-punishers/non-helpers (H14.1-H20.2) were registered as exploratory, as I did not know whether they would achieve 95% power. Results that did not meet the power requirements (H14.2a, H14.2b, H15.1, H19.2 and H20.1) are interpreted as suggestive, pending confirmation in future research.

Refer to Figure 6.1 for a visualisation of the hypotheses.

The accepted Stage 1 protocol for this Registered Report, which constitutes the full preregistration of all experiments in this chapter, is available at <https://doi.org/10.6084/m9.figshare.24559462.v1>.

## **Figure 6.1**

*Hypotheses for Study 1 (personal cost deliberation) and Study 2 (target impact deliberation).*



*Note.* In both studies, participants made an uncalculated or a calculated helping/punishing decision. Lower boxes indicate my expectations regarding whether a calculated or uncalculated decision is associated with a comparatively higher level of trust and trustworthiness. Green text and plus signs indicate my expectations of increased trust and trustworthiness, while red text and minus signs indicate my expectations of decreased trust and trustworthiness. This figure demonstrates the expectation that uncalculated helping will signal trustworthiness in the same way across studies, while uncalculated punishment will be similar to uncalculated helping when personal cost is deliberated, but differ when the impact on targets is considered.

## **6.2 Methods**

### **6.2.1 Ethics information**

The research complies with all relevant ethical regulations. The study was approved by the UCL Ethics Board (Project ID: *ICN-NH-PWB-7-1-23A*). Informed consent was obtained from all participants. Although ‘Player 3’/ ‘The Sender’ (Player A) and ‘The Receiver’ (Player B) really did exist and participants’ decisions really did influence their own payoff and that of fellow participants, ‘Player 1’ (the cheater in punishing contexts or the recipient in helping contexts) and ‘Player 2’ (the violator) did not actually exist. Therefore, participants were fully debriefed after the study, and only those who previously indicated they are willing to take part in studies involving deception were invited to the study. Participants were compensated at an hourly rate of £9.

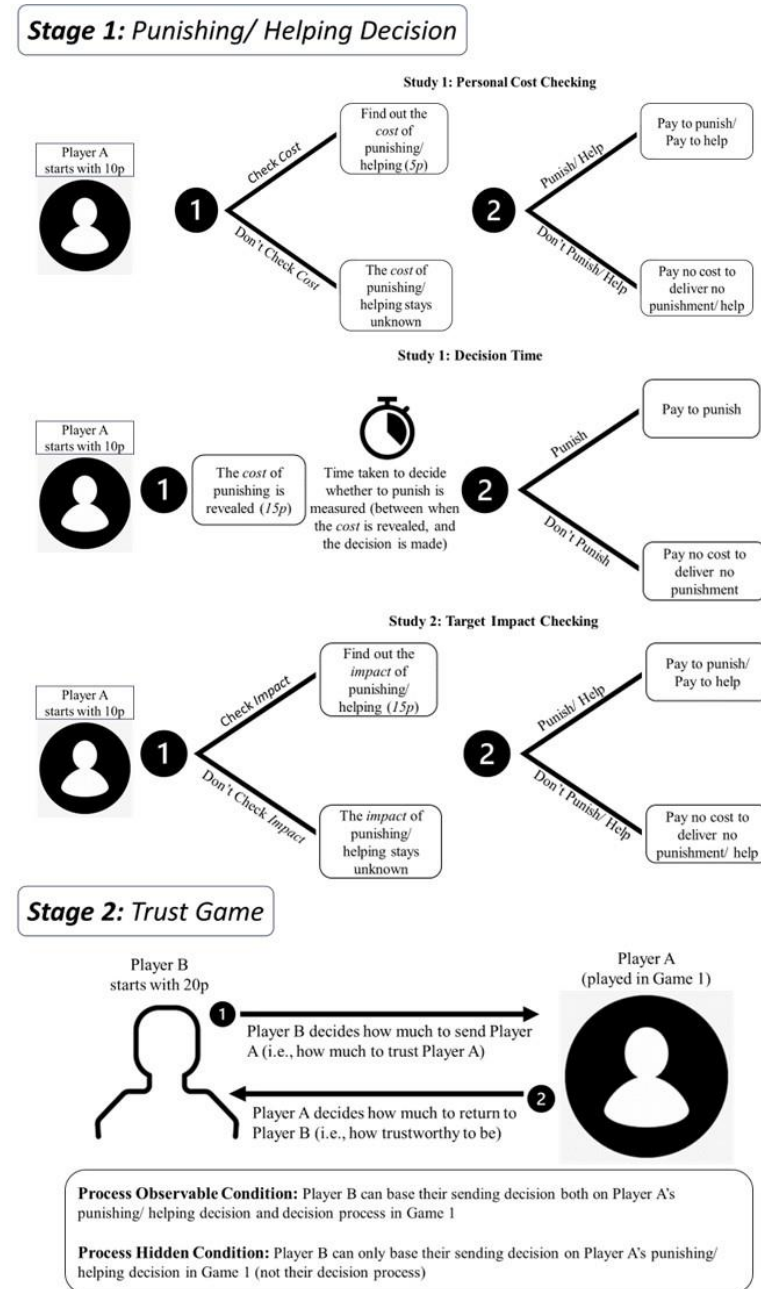
### **6.2.2 Design**

I conducted experiments of highly similar designs to investigate different operationalisations of each calculating behaviour: (i) checking the cost or impact of punishment or helping, and (ii) punishing cost decision time. Each experiment recruited separate sets of both Players A and Players B, and had two conditions: decision process hidden or decision process observable, with a between-subjects design. The studies were built in Qualtrics ([www.qualtrics.com](http://www.qualtrics.com)) and consisted of two-stage, incentivized, anonymous economic games (see Figure 6.2 for a visualisation of the study design). Players A made decisions during both games, whilst Players B only made decisions during the second game. Prior to making any decisions, Players A and Players B read the instructions for both games to ensure that they could make informed decisions and comprehension could be assessed.

The following describes the procedure for Study 1 in a punishing context. I outline any differences for the helping context below (see Appendix E2 for a selection of study materials).

**Figure 6.2**

*Illustration of the two-stage experimental design investigating uncalculated punishment and helping in Studies 1 and 2, for both checking behaviour and decision speed.*



*Note.* In Game 1 Player A could pay a cost to punish/help another player. Player A decided (i) whether to make their decision in a calculated or uncalculated way (operationalised via their cost-checking (Experiments 1.1 & 1.2a) or impact-checking (Experiments 2.1 & 2.2) decisions, or their decision time (Experiment 1.2b), and (ii) whether to punish/help. In Game 2, Player B decided how much to send Player A (indicating how much they trust Player A), and Player A decided how much to return to Player B (indicating how trustworthy Player A is). There were



two conditions in all experiments: in the process observable condition Player B could base their sending decisions both on Player A's decision process (i.e., their checking decision or decision time) as well as Player A's punishing/helping decision, whilst in the process hidden condition Player B could only make their decisions based on Player A's punishing/helping decision.

### **Stage 1**

Stage 1 is a standard Trust Game that was observed by both Player A and Player B. The outcome of a Trust Game played between 'Player 1' and 'Player 2' (who do not actually exist) were presented to Player A and B. Participants were told that Player 1 started with a £0.10 endowment and could choose how much of this endowment to send to Player 2, who started the trust game with no money. Participants were told that the amount entrusted was tripled by the experimenter and that Player 2 could then choose how much to return to Player 1. Participants were told that Player 1 and Player 2 already made their decisions: Player 1 sent their entire endowment (£0.10) to Player 2, who returned nothing, now leaving Player 1 with no money and Player 2 with £0.30.

Player A started with a £0.10 endowment and could then choose whether to use some of their endowment to 'punish' Player 2. Participants did not know exactly how much it would cost to punish Player 2, except that it would be somewhere between £0.00 and £0.10. Punishing always removed £0.15 from Player 2's bonus. I chose this amount as it leaves Player 2 with a bonus of £0.15, which would have been the 'fair' amount for Player 2 to receive in the trust game. Moreover, Player 2 still finishes Stage 1 with more than Player A, thereby reducing potential that punishment is perceived as stemming from Player A wanting to have more than Player 2.

Players A additionally decided whether to make their punishing decision in a calculating or an uncalculating manner. The way this was operationalised varied between participants across experiments (Table 6.1).

**Cost-checking:** Player A chose whether to check the cost of punishing Player 2 before making their punishing decision. The cost of punishment was always £0.05.

**Decision time:** Player A was told that the cost of punishing is £0.05, after which they immediately made their punishing decision on the same page. The time spent on this page was recorded to assess their decision time.

## **Stage 2**

In Stage 2, Players A and B participated in a trust game with each other. The trust games ‘rules’ were the same as in Stage 1. Player B started with a bonus of £0.10 and had to choose how much of their endowment (£0.00 - £0.10) to send to Player A. Any amount sent was tripled. Player A then chose what percentage of the endowment to return to Player B.

In the ‘process observable’ condition, Players B could condition their sending decision both on (i) Player A’s Game 1 decision process (whether Player A chose to reveal the cost of punishing in Experiment 1.2a, or Player A’s fast/slow decision time in Experiment 1.2b) and (ii) Player A’s Game 1 decision (whether Player A punished Player 2). In the ‘process hidden’ condition, Players B could only condition their sending decision on Player A’s Game 1 decision (whether they chose to punish Player 2). I employed the strategy method for both players: Players B decided how much to send to a Player A who engaged in all possible combinations of punishing decisions and (depending on the condition) processes, without knowing what Player A did. Similarly, Players A decided what percentage of the amount they received from Player B to return, without knowing how much Player B sent. Participants were told that the choice that matched the decision the other player made would determine their payoff.

All participants were asked several comprehension questions, primarily to assess their understanding of the incentive structure of both games. In addition to Player A decision time, I recorded the time spent on the three pages with comprehension questions in Experiment 1.2b. These recordings serve as a control for general comprehension and reading speed.

After completing the data collection, I randomly matched pairs of Players A and Players B who participated in the same experiment and condition. The decisions participants made during the study then determined their bonus payments.

## **Differences across studies and contexts**

The procedure across punishing and helping contexts was identical, except that instead of choosing whether to punish Player 2, Players A decided whether to use some of their endowment to help Player 1.

Whereas Study 1 explored how participants respond to an unknown personal cost of helping or punishing, Study 2 explored how participants respond to the unknown impact of helping or punishing another. In Study 2 participants were told that the personal cost of punishment (or helping) is £0.05 but they did not know exactly how much punishing would remove from Player

2 (or helping would benefit Player 1), except that it would be somewhere between £0.01 and £0.30. The maximum impact of £0.30 in Study 2 is equivalent to the maximum personal cost of £0.10 in Study 1: £0.10 was the entire endowment of the helper/punisher in Study 1, whereas £0.30 was the entire endowment of the target (in the punishment condition) in Study 2. Participants were informed that the minimum potential impact of helping/punishing a target is £0.01 because £0.00 would indicate no punishment or no help. When the impact was revealed, punishing still removed £0.15 from the target, and helping still delivered £0.15 to the target, just as in Study 1.

The procedure for Stage 2 only differed in that the Stage 1 procedure influenced what decisions and decision processes Players B could condition their sending decision on (i.e., whether it was a punishing or helping decisions and whether it centred around personal cost checking or impact checking).

### 6.2.3 Power Analysis

I conducted the power calculation in R (R Core Team, 2020) using the package ‘pwr’ (Champely, 2020) with the ‘pwr.f2.test’ function. I used a power of 0.95 with a 0.05 significance level and a one numerator degree of freedom ( $u$ , the number of coefficients in the model without the intercept). While estimating the required sample size, I referred to Jordan et al.'s (2016) supplementary materials for effect sizes, although specific effect sizes were not explicitly mentioned. I acknowledge that the available coefficients in their supplementary materials vary considerably, but generally produced small to medium effect sizes. As their study closely matches the present experimental design, procedure, and research questions, I used an effect size of  $f^2 = 0.02$  in my power analysis. According to Cohen's (1988) guidelines,  $f^2 \geq 0.02$  represents a small effect size (Selya et al., 2012). Because my main interests focussed on third-party punishment rather than helping, I expected to find similar or smaller effect sizes. Nevertheless, I must acknowledge that the choice of  $f^2 = 0.02$  might be considered a heuristic approximation rather than a precise estimation based on a formal inspection of Jordan et al.'s results. Based on the model with the highest number of predictors (as  $n = v + p$ , with  $p$  being the number of predictors including the intercept and  $v$  the degrees of freedom for the denominator), a sample size of 653 would be needed. However, as each of the models upon which this calculation was based involves either a Player A or a Player B participant, taking part in one of five experiments, in either the process hidden or in the process observable condition, a sample size of 13,060 (i.e., 1,306 Player A - Player B pairs per experiment) was

needed. As I could not predict how many participants would decide to punish/ help or not punish/ help, it was not possible to ascertain before data collection whether 95% power would be achieved for all analyses. Results that did not meet the power requirements are therefore interpreted as suggestive, pending confirmation in future research.

## **6.2.4 Participants**

I ran the experiments on Prolific (<https://prolific.co/>). Participants were invited to take part if they previously indicated on Prolific that they (i) are aged 18 or above, (ii) are from the UK, so that the currency specifications are familiar, (iii) are fluent in English, (iv) have the maximum approval rate of 100, and (v) selected “Yes, I would be comfortable to take part in such a study” to the question “Would you be happy to take part in a study where you are intentionally given inaccurate information about other participants and the study? You would be debriefed after the study”. To avoid participants taking part in more than one experiment, I launched the experiments in sequence, and allowed only new participants to take part. As preregistered, I lowered the approval rate to 97, in one-unit increments, as I did not reach enough participants with the maximum approval rate of 100. Median completion time for experiments ranged between six and seven minutes. Demographics were similar across experiments. Participants in Experiment 1 were aged between 18-80 years ( $M = 39.45$ ,  $SD = 12.51$ ) with 1381 women and 1207 men (17 identified as ‘other’ and 7 preferred not to say). In Experiment 2, participants were aged between 18-79 years ( $M = 39.16$ ,  $SD = 12.36$ ) with 1519 women and 1071 men (18 identified as ‘other’ and 7 preferred not to say). In Experiment 3, participants were predominantly women (1598 women, 990 men, 22 ‘other’, and 2 preferred not to say) and were aged between 18 and 91 years ( $M = 39.7$ ,  $SD = 12.65$ ). 63% of participants in Experiment 4 were women (1645 women, 949 men, 21 ‘other’, 2 preferred not to say) with a mean age of 37 years ( $SD = 11.79$ ; range: 18-80 years). In Experiment 5, the average age was 38 years ( $SD = 12.45$ ; range: 18-78 years), and 61% of participants were women (1581 women, 1002 men, 20 ‘other’, and 9 preferred not to say).

## **6.2.5 Data Exclusion**

I used the “force response” feature (which does not let participants proceed to the next page until they have provided a response) in Qualtrics to ensure that I did not receive incomplete responses. As in Jordan et al. (2016), responses by participants who failed more than one

attention check were still included in the analyses. However, I re-ran the same analyses excluding those who failed more than one attention check, and reported this version when it lead to significant differences in results. Any duplicate responses were removed.

### **6.3 Analysis Plan**

All analyses were conducted in R, and all analytical decisions for my hypotheses were independent of each other. For hypotheses in which I predicted cost-checking or impact-checking decisions (binary variables), I ran a logistic regression. For all other hypotheses I used linear regressions, as they predict decision speed, sending decisions, or returning decisions (continuous variables). Decision speed was a continuous variable when returning decisions were predicted, but a dummy variable (median split of relatively fast or slow) was used when predicting sending decisions. For ease of interpretation the measure for endowment sent was transformed from an absolute value (pence) sent to the percentage of endowment sent. The return measure was not transformed, as Players A already indicated what percentage, rather than what absolute value, they wished to return. In instances where Players B made sending decisions based on Player A's decision process, analyses were restricted to the process observable condition, as Players B did not know Player A decision processes in the process hidden condition. Due to Players B making multiple sending decisions in each of these analyses (based on the possible decisions made by Player A during the first stage), each sending decision was treated as an observation and robust standard errors were clustered on participant ID to account for the non-independence of repeated observations from the same participant (i.e., two observations per participant for decision process hidden, and four observations for decision process observable conditions). To accomplish this, I utilized the `lmtest` package (Zeileis & Hothorn, 2002), employing the functions `coeftest()` with the argument `vcov = vcovCL` to specify the use of the sandwich estimator, and `coefci()`. As data was collected on Player A's decision process in both conditions (even when Player B could not observe it), data from both the observable and hidden condition were used for analyses of returning decisions. As variance in decision time could also be caused by general comprehension ability or reading speed, rather than solely the time taken to reach a punishing decision, I included a control for general comprehension and reading speed when Player A decision time is an independent variable. General comprehension and reading speed was operationalised as the natural log transformed sum of time spent on the three comprehension question pages. As decision time was heavily skewed, punishing decision time was natural-log transformed, according to preregistrations. All

reported coefficients are unstandardised. For more detail of individual analysis methods for the hypotheses, see Appendix E1.

### 6.3.1 Bayesian Analyses

In addition to the frequentist analyses, I conducted equivalent Bayesian analyses to assess the evidence for each hypothesis compared to the null hypothesis. These analyses primarily served to judge whether frequentist null-findings were due to a need for more data or the absence of an effect. I used the BayesFactor package (Morey & Rouder, 2022) with the `lmBF` function for linear regressions with return decisions as the response variable. For the logistic regressions and linear regressions with sending decisions as the response variable, I used the `brm()`, `bridge_sampler()` and `bayes_factor()` functions from the `brms` package (Bürkner, 2017), which is better suited to handle the repeated observations from Player Bs. For analyses with repeat observations, I fit mixed models in `brm()` with participant ID as random effect.

I constructed effect priors that are zero-centred t-distribution priors with 4 degrees of freedom. The prior width was designed such that only one-third of the prior mass on each side of zero is larger than the desired effect (i.e., the relevant coefficient observed in Jordan et al., 2016). Specifically, for any desired effect, one-third of the prior mass on each side of zero was more extreme than the absolute value of the desired effect. The total prior mass smaller than the desired effect was calculated as  $0.5 + 0.5 * 2/3 = 0.83333$  (e.g. assuming an effect of 5.6 would lead to an effect prior with scale of 5.09). To achieve this, I used the below R code to calculate the scale of the prior width for a given desired effect (e.g. 5.6):

```
desiredEffect <- 5.6
myt <- function(x) { abs(extraDistr::qlst(0.5 + 0.5 * 2/3 , df = 4,
mu = 0, sigma = x) - desiredEffect) }
calc_scale <- optimize(myt, interval = c(0, 20))
prior_width_scale <- calc_scale$minimum
```

The defined function "myt" calculates the absolute difference between the desired effect size (in this example 5.6) and the quantile of the t-distribution with 4 degrees of freedom and zero mean, corresponding to the prior mass of  $0.5 + 0.5*2/3$ . The "optimize" function in R then finds the value of the scale parameter that minimizes the absolute difference between the observed

effect size and the quantile of the t-distribution. The resulting `prior_width_scale` value is what I used as the width of my prior distributions.

I chose this specification because previous research from Jordan et al. (2016) indicates effect sizes may generally be small, and as these studies investigated punishment as well as helping, effect sizes in punishing contexts may be smaller still. However, I still allowed for the possibility that I could sometimes find larger effects.

All other priors were weakly informative, using a zero-centred t-prior with 4 degrees of freedom and a scale of 10. I chose these weakly informative priors to allow for some flexibility in the effect size estimates while still constraining them to reasonable values. The choice of 4 degrees of freedom and a scale of 10 reflects my prior belief that the effect size was unlikely to be very large, but may occasionally have been larger than expected.

To ensure that my prior on the effect size was appropriate, I set `rscalefixed = 0.5` when using the BayesFactor package, as this parameter specifies the scale of the prior distribution for the fixed effects. A value of 0.5 is the smallest recommended prior on the effect size (Singmann et al., 2023).

For hypotheses investigating the effect of deliberation on trust and trustworthiness, based on whether it is measured through cost checking or decision time, or takes place in the context of helping or punishing, I could not directly rely on equivalent coefficients from previous research to set priors as I did above. However, as I expected small effects, I set `rscalefixed = 0.5` here as well, and to be conservative used the smallest interaction effect found in Jordan et al. (2016) to calculate the prior scale for analyses in which sending decisions are the response variable.

I also conducted sensitivity analyses for each Bayes factor test by conducting two additional analyses: one with a prior scale of 0.5 times the original value and one with a prior scale of 1.5 times the original value. As none changed the direction of the Bayes factor, these results can be found in Appendix E3.

In evaluating the strength of evidence for or against the alternative hypothesis compared to the null hypothesis, I used common decision heuristics (Jeffreys, 1939; Lee & Wagenmakers, 2014) and considered Bayes factors of 3 as weak evidence in favour of the alternative hypothesis, and Bayes factors of one-third as weak evidence in favour of the null hypothesis over the alternative hypothesis. Bayes factors of 10 or more were considered substantial evidence for the alternative hypothesis. Conversely, Bayes factors of one-tenth or less were considered substantial evidence for the null hypothesis. In cases where the Bayes factor fell between these thresholds, I

concluded that the data provided no strong evidence for either the alternative or the null hypothesis and that more data were needed to draw a conclusive inference.

## **6.4 Deviations from the Preregistration**

Due to some participants starting but not finishing the experiment, some condition cells were unbalanced. I therefore recruited an additional 13 participants (five each in Experiments 1 and 4, and three in Experiment 2) to ensure that each condition reached the preregistered number of participants, bringing the total sample size to 13073 rather than 13060. Originally, I planned to run Bayesian analyses for H9.2b with `lmBF()`. However, as the function currently does not allow for models containing both continuous and categorical predictors, those models were fit with `brm()` as specified in the Bayesian Analysis section instead. Lastly, to maintain consistency with my registered analyses, I incorporated additional analyses centring around non-action in the comparison between help and punishment, ensuring comprehensive coverage and completeness across all hypotheses.



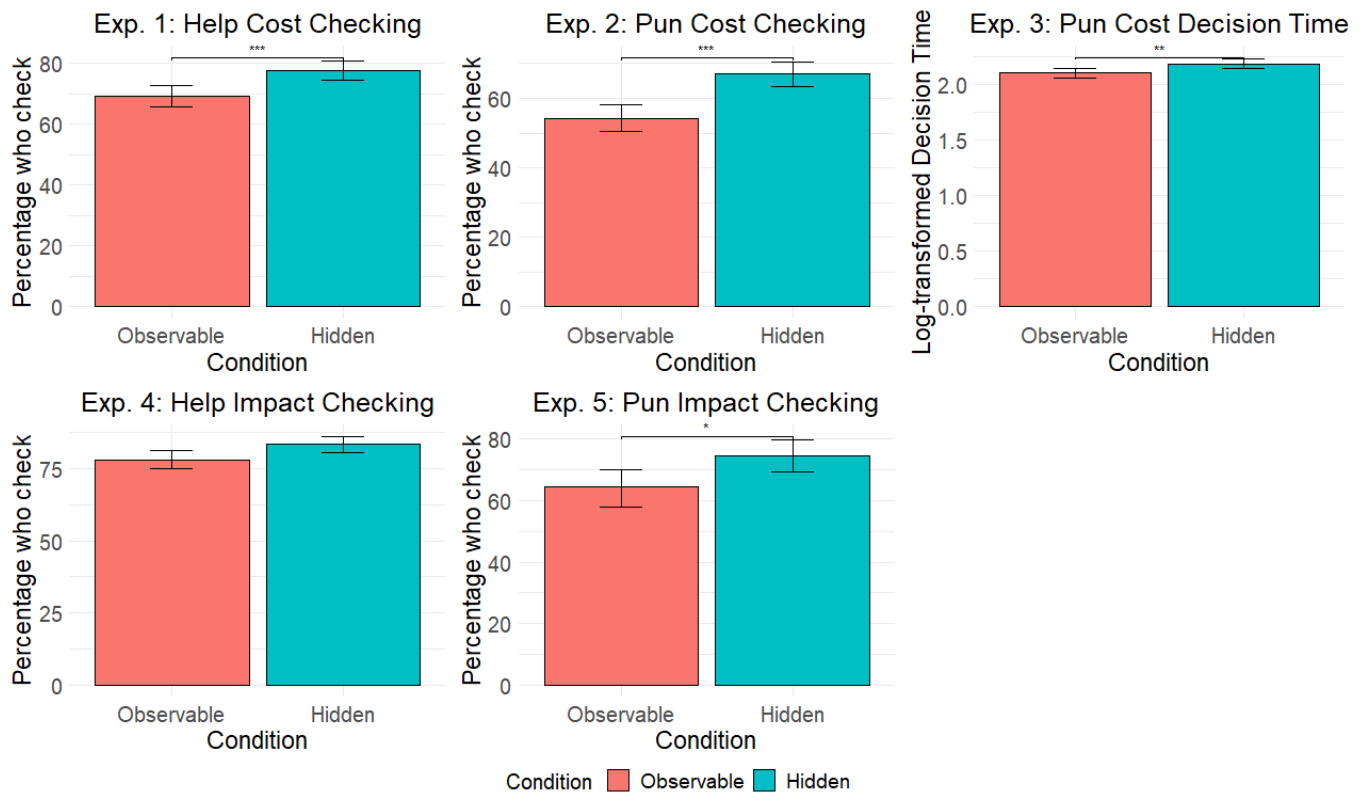
## 6.5 Results

### 6.5.1 Is deliberation around help and punishment used as a signal of trustworthiness?

A main prediction was that people would strategically adjust their decision-making process to gain reputational benefits. When considering the personal cost of helping (Exp. 1) and punishment (Exp. 2 & Exp. 3), as well as when considering the impact of helping (Exp. 4), I expected Players A to be more likely to make uncalculating decisions (by deciding quickly or without checking the personal cost or target impact) when their decision process was observed by others (and could therefore potentially confer reputation benefits), compared to when their decision process was hidden. I predicted the opposite when participants considered the impact of punishment on a target (Exp. 5): as punishment is a harmful act, I expected Players A to make calculating decisions (by checking the impact that punishment would have on the target) when their decision process was observed. See Figure 6.3 for a visualisation of results and Table 6.1 for the nomenclature of experiments.

#### Figure 6.3

*Decision processes (uncalculating versus calculating) across all five experiments, in the decision process observable and hidden conditions.*



*Note.* Checking (versus not checking) the personal cost or target impact (Exp. 1, 2, 4 & 5), as well as taking a long (versus a short) time to decide (Exp. 3), reflect calculated decision making. Error Bars indicate 95% confidence intervals (CIs). Due to changes in significance levels, bar charts for Exp. 4 and Exp. 5 only include participants with excellent comprehension ( $n = 1311$  for Exp. 1,  $n = 1309$  for Exp. 2,  $n = 1306$  for Exp. 3,  $n = 534$  for Exp. 4,  $n = 535$  for Exp. 5). Differences are significant for all but Exp. 4 (help impact checking).

In line with predictions, participants were around half as likely (odds ratio (OR) = 0.64, 95% CI [0.50, 0.82]) to check the personal cost of helping when their decision process was observable than when it was hidden (Exp. 1, H1.1). Specifically, 78% of Players A checked the cost of helping in the decision process hidden condition, whereas only 69% did so in the decision process observable condition ( $b = -0.44$ , 95% CI [-0.69, -0.19],  $p < .001$ , BF = 71.29).

Similarly, the odds of checking the personal cost of punishing decreased by around 41% (OR = 0.59, 95% CI [0.47, 0.73]) when the participants' decision process was observable compared to hidden (Exp. 2, H1.2a). Specifically, 67% of Players A checked the cost of punishing in the decision process hidden condition, whereas only 54% did so in the decision process observable condition ( $b = -0.53$ , 95% CI [-0.76, -0.31],  $p < .0001$ , BF = 6972.38).

Players A also made faster punishing decisions when their decision process was observable ( $M = 2.11$  log-seconds, 95% CI [2.06, 2.15]) compared to hidden ( $M = 2.19$  log-seconds, 95% CI [2.14, 2.23]), with a 7.86% decrease in decision time in the observable condition ( $b = -0.08$ , 95% CI [-0.14, -0.02],  $t(1304) = -2.68$ ,  $p = .007$ ,  $BF = 2.16$ ) (Exp. 3, H1.2b).

In Study 2 (checking the impact of helping or punishing on a target), results were not as clear-cut. As predicted, participants were significantly more likely to check the impact of helping when their decision process was hidden (83%) compared to observable (78%) ( $b = -0.34$ , 95% CI [-0.62, -0.06],  $p = .016$ ,  $BF = 3.86$ ;  $OR = 0.71$ , 95% CI [0.54, 0.94]). Nevertheless, when only those who correctly responded to at least 7 out of 8 comprehension questions were included (hereafter referred to as participants with excellent comprehension), results were in the same direction but no longer significant (81% checked the impact when their decision process was hidden compared to 77% when it was observable;  $b = -0.24$ , 95% CI [-0.66, 0.18],  $p = .26$ ,  $BF = 0.67$ ;  $OR = 0.78$  [0.51, 1.19]) (Exp. 4, H5.1).

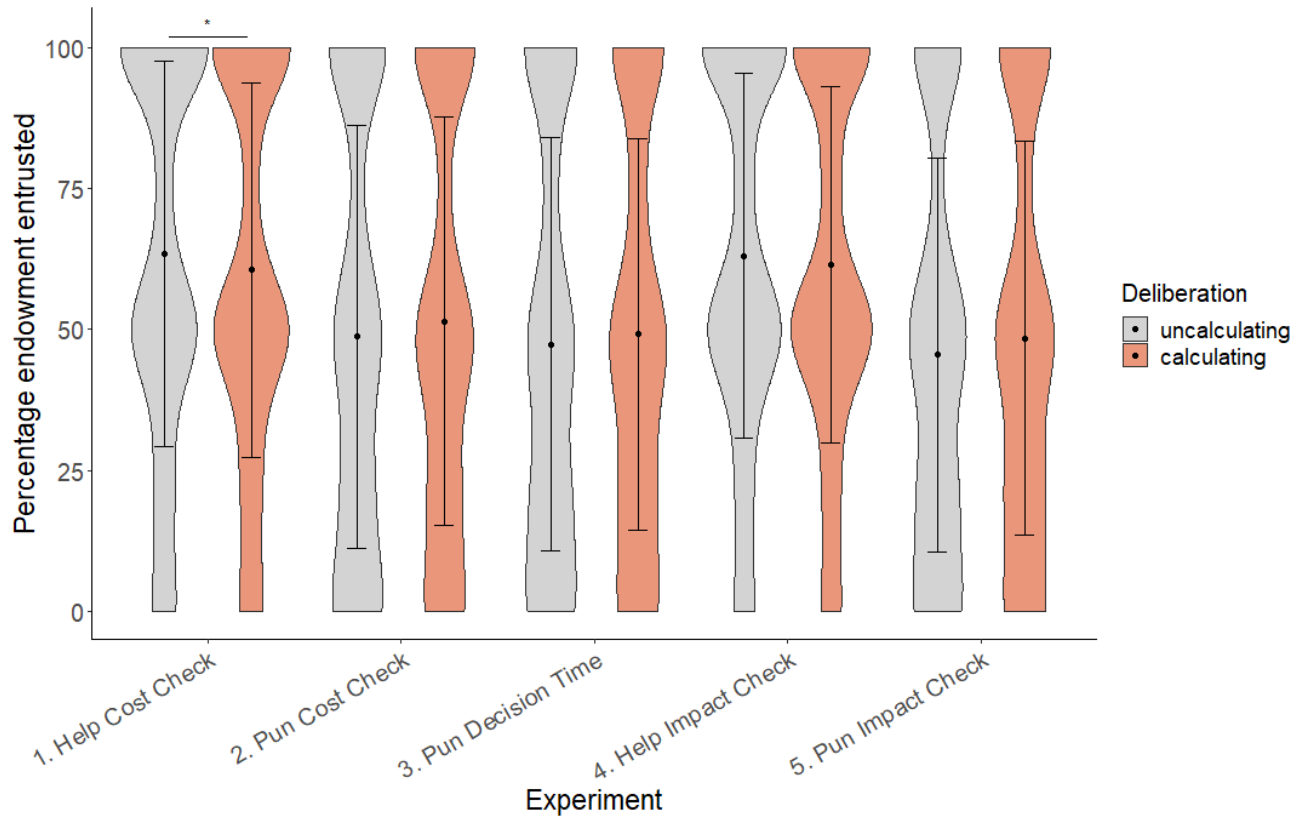
I expected that participants would be more likely to check the impact of punishment when their decision process was observed (Exp. 5, H5.2), but this prediction was not supported by the data ( $b = -0.17$ , 95% CI [-0.41, 0.07],  $p = .16$ ,  $BF = 0.60$ ;  $OR = 0.84$  [0.66, 1.07]). Indeed, when only participants with excellent comprehension were included, I found the opposite: 75% of Players A checked the impact of punishing in the decision process hidden condition, whereas only 64% did so in the decision process observable condition ( $b = -0.49$ , 95% CI [-0.86, -0.11],  $p = .011$ ,  $BF = 5.89$ ;  $OR = 0.62$ , 95% CI [0.42, 0.89]).

### **6.5.2 Is deliberation around help and punishment *perceived* as a signal of trustworthiness?**

Next, I explore how helping and punishment decisions were interpreted by observers. I expected uncalculated help and punishment in the context of personal cost deliberation, as well as uncalculated help and calculated punishment in the context of target impact deliberation, to confer reputational benefits. Specifically, I expected observers to send helpers and punishers a higher percentage of their endowment in those situations, which I interpret as higher trust. See Figure 6.4 for a visualisation of results.

**Figure 6.4**

*Percentage of endowment sent to uncalculating and calculating helpers (Exp. 1 & Exp. 4) and punishers (Exp. 2, Exp. 3 & Exp. 5) by observers.*



*Note.* Checking (versus not checking) the personal cost or target impact, as well as taking a long (versus a short) time to decide reflect calculated decision making. The width of the violins indicate the distribution of observations, error bars indicate 95% CI, dots represent the mean. Due to changes in significance levels, Exp. 1 (help cost checking) only includes participants with excellent comprehension ( $n = 612$  for Exp. 1,  $n = 1306$  for Exp. 2,  $n = 1306$  for Exp. 3,  $n = 1306$  for Exp. 4,  $n = 1306$  for Exp. 5). Differences are significant for Exp. 1 (help cost checking).

In line with predictions, observers sent a higher proportion of their endowment to helpers who did not check the personal cost of helping ( $M = 63.40\%$ ,  $SD = 34.24$ ) than to helpers who checked the cost ( $M = 60.57\%$ ,  $SD = 33.33$ ) ( $b = -2.83$ , 95% CI  $[-6.50, 0.82]$ ,  $t(1304) = -1.52$ ,  $p = .13$ ,  $BF = 6.00$ ) (Exp. 1, H2.1). Although non-significant, the Bayes Factor indicates support for H2.1, and when only participants with excellent comprehension were included in the analysis, this finding was statistically significant (uncalculating helper:  $68.27\%$  ( $SD = 33.89$ );

calculating helper: 62.65% ( $SD = 33.73$ );  $b = -5.62$ , 95% CI  $[-10.99, -0.26]$ ,  $t(610) = -2.06$ ,  $p = .04$ ,  $BF = 8.97$ ).

My predictions that observers would send more money to punishers who made uncalculating decisions (when considering personal costs) were not supported. If anything, observers entrusted a higher proportion of their endowment to punishers who checked the personal cost of punishment ( $M = 51.49\%$ ,  $SD = 36.26$ ) than to those who did not ( $M = 48.76\%$ ,  $SD = 37.62$ ) (Exp. 2, H2.2a). However, this difference was statistically non-significant ( $b = 2.73$ , 95% CI  $[-1.28, 6.74]$ ,  $t(1304) = 1.33$ ,  $p = 0.18$ ,  $BF = 4.78$ ). When calculating behaviour was operationalised in terms of decision time, observers sent more to relatively slow (more calculating) punishers ( $M = 49.17\%$ ,  $SD = 34.75$ ) than to relatively fast punishers ( $M = 47.40\%$ ,  $SD = 36.80$ ) (Exp. 3, H2.2b). Again, this result was not statistically significant ( $b = -1.78$ , 95% CI  $[-5.66, 2.11]$ ,  $t(1304) = -0.90$ ,  $p = .37$ ,  $BF = 0.39$ ).

I expected that helpers who did not check the impact of helping behaviour would be trusted more by observers. Although observers sent a higher percentage of their endowment to helpers who did not check the impact ( $M = 63.12\%$ ,  $SD = 32.39$ ) than to those who did ( $M = 61.53\%$ ,  $SD = 31.68$ ) (Exp. 4, H6.1), this result was statistically non-significant ( $b = -1.60$ , 95% CI  $[-5.07, 1.88]$ ,  $t(1304) = -0.90$ ,  $p = .37$ ,  $BF = 1.53$ ).

Conversely, I expected that impact-checking punishers would be trusted more by observers. Although observers did directionally send more of their endowment to punishers who checked the impact of punishing ( $M = 48.45\%$ ,  $SD = 34.93$ ) than to punishers who did not ( $M = 45.54\%$ ,  $SD = 35.01$ ) (Exp. 5, H6.2), this difference was also statistically non-significant ( $b = 2.91$ , 95% CI  $[-0.89, 6.71]$ ,  $t(1304) = 1.50$ ,  $p = .13$ ,  $BF = 4.31$ ).

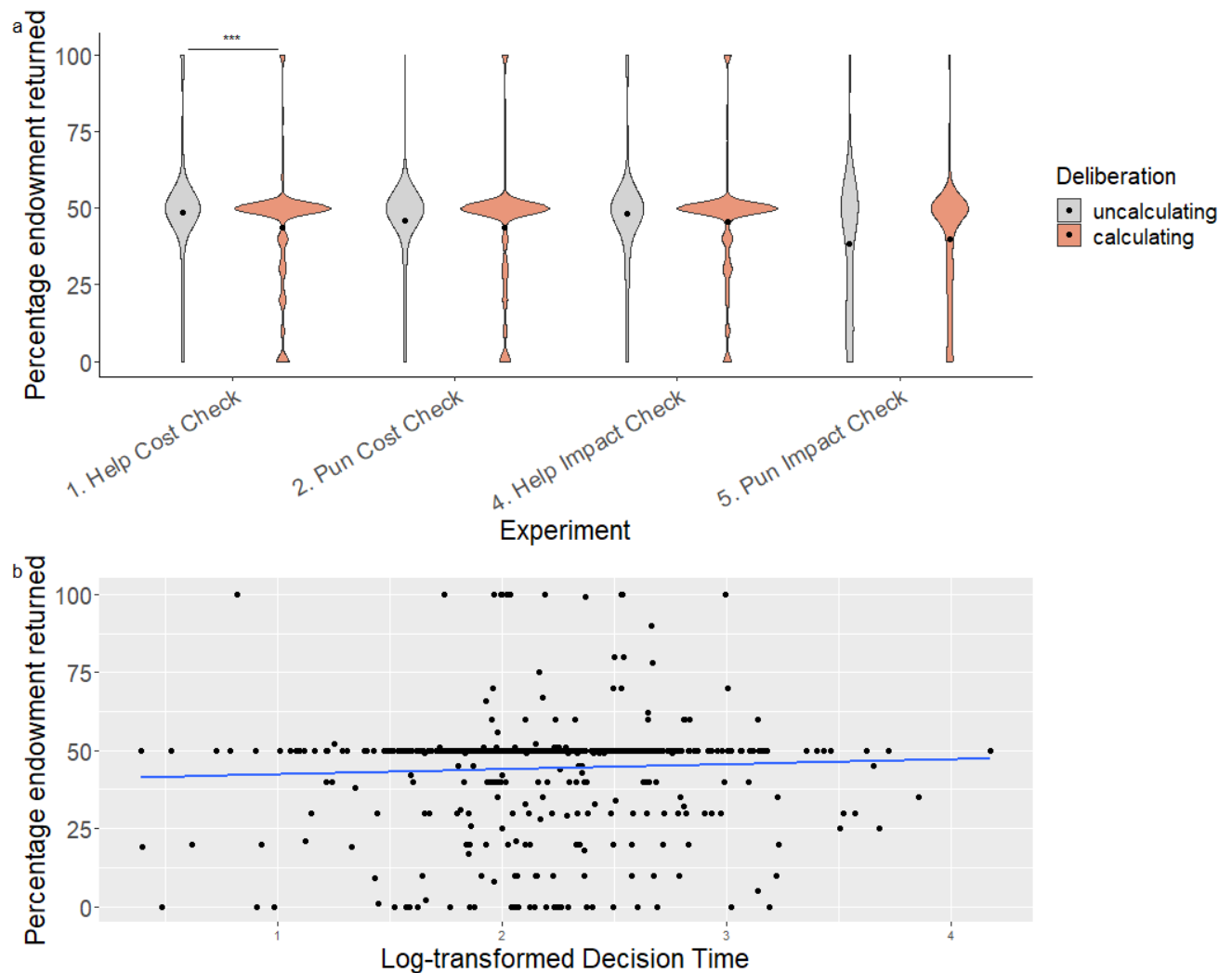
### **6.5.3 Is deliberation around help and punishment *actually* a signal of trustworthiness?**

Next, I asked whether calculated/uncalculated help and punishment decisions reliably signalled a person's trustworthiness (Figure 6.5). I expected uncalculating helpers in both the personal cost (Exp. 1) and impact checking context (Exp. 4) to be more trustworthy than calculating helpers. Indeed, helpers who did not check the personal cost of helping returned significantly more of the endowment they were sent by observers ( $M = 48.74\%$ ,  $SD = 19.76$ ) than helpers who did check the personal cost ( $M = 43.54\%$ ,  $SD = 19.34$ ) ( $b = -5.21$ , 95% CI  $[-7.86, -2.55]$ ,

$t(1099) = -3.85, p < .001, BF = 97.33$ ) (Exp. 1, H14.1). Similarly, helpers who did not check the impact of helping ( $M = 48.23\%$ ,  $SD = 19.04$ ) returned a higher percentage in the Trust Game than helpers who checked the impact of helping ( $M = 45.40\%$ ,  $SD = 18.57$ ), but this effect was statistically non-significant ( $b = -2.83$ , 95% CI  $[-5.67, 0.004]$ ,  $t(1138) = -1.96, p = .05$ ,  $BF = 0.44$ ) (Exp. 4, H19.1).

**Figure 6.5**

*Percentage of endowment returned to observers by uncalculating and calculating helpers (Exp. 1 & Exp. 4) and punishers (Exp. 2, Exp. 3 & Exp. 5) in the Trust Game.*



*Note.* Checking (versus not checking) the personal cost or target impact (Exp. 1, 2, 4 & 5; Panel a), as well as taking a longer time to decide (Exp. 3; Panel b) reflect calculated decision making. In Panel (a) the width of the violins indicate the distribution of observations, error bars indicate 95% CI, dots represent the mean. Panel (b) shows a scatterplot with regression line. Differences

are significant for Exp. 1 (help cost checking). Participant numbers vary across experiments ( $n = 1101$  for Exp. 1,  $n = 508$  for Exp. 2,  $n = 515$  for Exp. 3,  $n = 1140$  for Exp. 4,  $n = 410$  for Exp. 5).

I expected that punishers who made an uncalculating versus calculating decision in the context of personal cost (Exp. 2 and 3) would be more trustworthy. Conversely, for impact consideration, I predicted that punishers who made calculating decisions would be more trustworthy than punishers who made uncalculating decisions (Exp. 5). Results do not support these predictions. Although punishers who did not check the personal cost of punishing ( $M = 46.11\%$ ,  $SD = 15.48$ ) returned more of the endowment they were sent by observers than punishers who did check the personal cost ( $M = 43.78\%$ ,  $SD = 20.01$ ), this difference was not statistically significant ( $b = -2.33$ , 95% CI  $[-6.76, 2.09]$ ,  $t(506) = -1.04$ ,  $p = .30$ ,  $BF = 0.17$ ) (Exp. 2, H14.2a). Conversely, when uncalculating decisions were instead operationalised as decision time, punishers who made slower (more calculating) decisions returned a slightly higher percentage than those who made faster (uncalculating) punishing decisions ( $b = 0.98$ , 95% CI  $[-1.94, 3.90]$ ,  $t(513) = 0.66$ ,  $p = .51$ ,  $BF = 0.14$ ) (Exp. 3, H14.2b). This difference was not significant. Punishers who did not check the impact of punishing returned a lower percentage ( $M = 38.48\%$ ,  $SD = 22.53$ ) than punishers who did check the impact of punishing on the target ( $M = 40.03\%$ ,  $SD = 19.67$ ),  $b = 1.55$ , 95% CI  $[-3.68, 6.78]$ ,  $t(408) = 0.58$ ,  $p = .56$ ,  $BF = 0.12$  (Exp. 5, H19.2). Although directionally in line with predictions, this difference too was non-significant.

It must be noted, that hypotheses H14.2a, H14.2b and H19.2 did not meet power requirements, therefore making their results suggestive, pending confirmation in future research. However, their Bayes Factor values do indicate support for the null hypotheses (see Appendix E3 for sensitivity analyses).

#### **6.5.4 Do trust and trustworthiness differ across experiments?**

I expected uncalculated decision-making to differentially influence trust and trustworthiness across the experiments (Exp. 1-3) of Study 1 (personal cost). Firstly, for punishment, I predicted that deliberation would have a stronger influence on trust and trustworthiness when calculating behaviour was operationalised as cost-checking (Exp. 2) compared to slow decision time (Exp.

3). In addition, I expected deliberation to have a stronger effect on trust and trustworthiness in the context of helping compared to punishing (Exp. 1 vs Exp. 2).

However, the effect of calculated versus uncalculated punishment on trust was not stronger for cost checking than decision time ( $b = 0.95$ , 95% CI  $[-1.99, 3.88]$ ,  $t(2608) = 0.33$ ,  $p = .74$ ,  $BF = 0.10$ ) (H3). The same was true for non-punishment ( $b = -0.55$ , 95% CI  $[-3.45, 2.35]$ ,  $t(2608) = -0.19$ ,  $p = .85$ ,  $BF = 0.75$ ) (H10). Similarly, the effect of calculated versus uncalculated punishment on trustworthiness was not stronger for cost checking than decision time,  $b = -3.44$ , 95% CI  $[-8.78, 1.90]$ ,  $t(1020) = -1.27$ ,  $p = .21$ ,  $BF = 0.24$  (H16). Again, the same was true for non-punishment:  $b = -3.80$ , 95% CI  $[-8.13, 0.54]$ ,  $t(1587) = -1.72$ ,  $p = .09$ ,  $BF = 0.33$  (H17).

Observers trusted helpers significantly more than they trusted punishers ( $b = 14.64$ , 95% CI  $[10.74, 18.54]$ ,  $t(2608) = 7.47$ ,  $p < .0001$ ). Moreover, trust was significantly influenced by the interaction between behaviour (helping versus punishing) and decision process (calculating versus uncalculating) ( $b = -5.56$ , 95% CI  $[-8.29, -2.83]$ ,  $t(2608) = -2.01$ ,  $p = .04$ ,  $BF = 140.10$ ) (H4). Specifically, uncalculating punishers were trusted the least ( $M = 48.76\%$ ,  $SD = 37.62$ ), followed by calculating punishers ( $M = 51.49\%$ ,  $SD = 36.26$ ), calculating helpers ( $M = 60.57\%$ ,  $SD = 33.33$ ), and uncalculating helpers ( $M = 63.40\%$ ,  $SD = 34.24$ ). Uncalculating helpers were trusted significantly more than uncalculating punishers ( $b = 14.64$ ,  $t(2608) = 7.47$ ,  $p < .0001$ ) and calculating helpers were trusted significantly more than calculating punishers ( $b = 9.08$ ,  $t(2608) = 4.64$ ,  $p < .0001$ ).

This interaction was no longer significant when excluding those who failed more than one comprehension check ( $b = -6.63$ , 95% CI  $[-10.66, -2.60]$ ,  $t(1162) = -1.54$ ,  $p = .12$ ,  $BF = 20.22$ ). Nevertheless, observers still trusted helpers significantly more than punishers ( $b = 13.00$ , 95% CI  $[6.87, 19.13]$ ,  $t(1605) = 4.28$ ,  $p < .0001$ ). Specifically, observers trusted uncalculating punishers the least ( $M = 55.27\%$ ,  $SD = 40.76$ ), increasing their levels of trust for calculating punishers ( $M = 56.28\%$ ,  $SD = 38.23$ ), calculating helpers ( $M = 62.65\%$ ,  $SD = 33.73$ ), and uncalculating helpers ( $M = 68.27\%$ ,  $SD = 33.89$ ).

There was also no evidence to suggest that the effect of calculated versus uncalculated decision-making on trustworthiness is stronger in helping compared to punishing contexts ( $b = -2.88$ , 95% CI  $[-8.04, 2.29]$ ,  $t(1605) = -1.09$ ,  $p = .28$ ,  $BF = 0.17$ ) (H18), and there was no significant difference in the trustworthiness of helpers and punishers ( $b = 2.63$ , 95% CI  $[-2.01, 7.27]$ ,  $t(1605) = 1.11$ ,  $p = .27$ ).



### **6.5.5 Is deliberation around decisions *not* to help or punish perceived as a signal of trustworthiness?**

Moreover, I had diverging expectations for how uncalculating decisions are perceived when those decisions result in inaction rather than helping or punishing. I predicted that observers would send more to calculating than uncalculating non-helpers/non-punishers in Experiments 1, 4 and 5, but more to uncalculating than calculating non-punishers when personal cost is being considered (Exp. 2 and Exp. 3). However, none of these analyses were statistically significant.

Directionally in line with predictions, observers sent more of their endowment to non-helpers who checked the cost of helping ( $M = 29.75\%$ ,  $SD = 33.70$ ) than to non-helpers who did not check the cost of helping ( $M = 28.81\%$ ,  $SD = 34.38$ ) ( $b = 0.95$ , 95% CI  $[-2.75, 4.64]$ ,  $t(1304) = 0.50$ ,  $p = .61$ ,  $BF = 0.68$ ) (Exp.1, H7.1). Conversely, and again in line with predictions, in Experiment 2 (H7.2a) observers sent directionally less of their endowment to non-punishers who checked the personal cost of punishing ( $M = 50.31\%$ ,  $SD = 35.38$ ) than to non-punishers who did not check the cost of punishing ( $M = 51.58\%$ ,  $SD = 37.83$ ) ( $b = -1.27$ , 95% CI  $[-5.25, 2.70]$ ,  $t(1304) = -0.63$ ,  $p = .53$ ,  $BF = 0.80$ ). However, Experiment 3 (H7.2b) found that observers sent more of their endowment to relatively slow (calculating) non-punishers ( $M = 49.36\%$ ,  $SD = 37.18$ ) than to relatively fast (uncalculating) non-punishers ( $M = 48.64\%$ ,  $SD = 35.97$ ) ( $b = 0.72$ , 95% CI  $[-3.25, 4.69]$ ,  $t(1304) = 0.36$ ,  $p = .72$ ,  $BF = 0.26$ ). In Experiment 4 (H11.1) observers were again in line with predictions and sent more of their endowment to non-helpers who checked the impact of helping ( $M = 32.43\%$ ,  $SD = 34.22$ ) than to non-helpers who did not ( $M = 29.71\%$ ,  $SD = 33.17$ ) ( $b = 2.73$ , 95% CI  $[-0.93, 6.38]$ ,  $t(1304) = 1.46$ ,  $p = .14$ ,  $BF = 5.05$ ). In Experiment 5 (H11.2) observers sent similar amounts of their endowment to non-punishers who checked the impact of punishing ( $M = 52.48\%$ ,  $SD = 34.09$ ) and to non-punishers who did not ( $M = 52.85\%$ ,  $SD = 34.68$ ) ( $b = -0.37$ , 95% CI  $[-4.10, 3.37]$ ,  $t(1304) = -0.19$ ,  $p = .85$ ,  $BF = 0.48$ ).

### **6.5.6 Is deliberation around decisions *not* to help or punish a signal of trustworthiness?**

Similarly to perceived trustworthiness, I also had diverging expectations for how uncalculating decisions predict the actual trustworthiness of non-helpers and non-punishers. Specifically, I predicted that calculating non-punishers in the context of impact checking (Exp. 5) and

calculating non-helpers in both the context of impact (Exp. 4) and cost checking (Exp. 1) would return more than uncalculating non-helpers/non-punishers. In contrast, I expected uncalculating non-punishers to return more than calculating non-punishers in context of personal cost deliberation (Exp. 2 & Exp. 3). All returning decisions for non-punishers and non-helpers were directionally in line with predictions.

In Experiment 1 (H15.1), non-helpers who checked the personal cost of helping returned more of their endowment ( $M = 21.29\%$ ,  $SD = 22.95$ ) than non-helpers who did not check the cost ( $M = 15.69\%$ ,  $SD = 22.88$ ) ( $b = 5.60$ , 95% CI  $[-1.00, 12.19]$ ,  $t(208) = 1.67$ ,  $p = .10$ ,  $BF = 0.56$ ). However, this difference was non-significant. As predicted, in Experiment 2 (H15.2a), non-punishers who did not check the cost of punishing ( $M = 42.51\%$ ,  $SD = 21.65$ ) returned significantly more of their endowment than non-punishers who checked the cost ( $M = 38.51\%$ ,  $SD = 22.42$ ) ( $b = -4.0$ , 95% CI  $[-7.06, -0.94]$ ,  $t(799) = -2.57$ ,  $p = .01$ ,  $BF = 1.98$ ). In Experiment 3 (H15.2b) uncalculating (faster) non-punishers again returned more of their endowment than calculating (slower) non-punishers, but this was not significant ( $b = -0.43$ , 95% CI  $[-3.33, 2.47]$ ,  $t(787) = -0.29$ ,  $p = .77$ ,  $BF = 0.01$ ). In Experiment 4 (H20.1) non-helpers who checked the impact of helping ( $M = 25.44\%$ ,  $SD = 23.84$ ) returned significantly more of their endowment than non-helpers who did not check the impact ( $M = 17.24\%$ ,  $SD = 26.56$ ) ( $b = 8.21$ , 95% CI  $[0.06, 16.35]$ ,  $t(169) = 1.99$ ,  $p = .048$ ,  $BF = 1.02$ ). However, the difference (calculating non-helper: 23.40% ( $SD = 24.55$ ), uncalculating non-helper: 16.00% ( $SD = 24.11$ )) was no longer statistically significant when only those with excellent comprehension were included ( $b = 7.40$ , 95% CI  $[-5.98, 20.78]$ ,  $t(58) = 1.11$ ,  $p = .27$ ,  $BF = 0.44$ ). Finally, in Experiment 5 (H20.2) both non-punishers who checked the impact of punishing ( $M = 38.39\%$ ,  $SD = 22.74$ ) and non-punishers who did not check the impact ( $M = 38.26\%$ ,  $SD = 24.68$ ) returned around 38% of their endowment ( $b = 0.13$ , 95% CI  $[-3.17, 3.43]$ ,  $t(822) = 0.08$ ,  $p = .94$ ,  $BF = 0.08$ ).

It must be noted that power requirements were not met for hypotheses H15.1 (Exp. 1) and H20.1 (Exp. 4), making those results suggestive, pending confirmation in future research.

### **6.5.7 Does deliberation have a stronger influence on perceived and actual trustworthiness when decisions result in helping or punishing versus inaction?**

Furthermore, for all experiments I predicted that the effect of uncalculating behaviour on trust and trustworthiness would be larger for action than inaction, meaning that deliberation would

more strongly influence sending and returning decisions when Player A decided to help/punish compared to when Player A decided *not* to help/punish.

However, for sending decisions this was not the case in Experiment 1 (H8.1;  $b = -3.78$ , 95% CI [-8.99, 1.42],  $t(2608) = -1.43$ ,  $p = .15$ , BF = 1.16), Experiment 2 (H8.2a;  $b = 4.0$ , 95% CI [-1.65, 9.64],  $t(2608) = 1.39$ ,  $p = .17$ , BF = 1.17), Experiment 3 (H8.2b;  $b = -2.50$ , 95% CI [-5.40, 0.40],  $t(2608) = -0.88$ ,  $p = .38$ , BF = 0.28), or Experiment 5 (H12.2;  $b = 3.28$ , 95% CI [-2.05, 8.60],  $t(2608) = 1.21$ ,  $p = .23$ , BF = 0.79). Yet, when only participants with excellent comprehension were included, there was a significant interaction between deliberation and helping decision in Experiment 4 (H12.1;  $b = -7.97$ , 95% CI [-15.54, -0.40],  $t(1060) = -2.06$ ,  $p = .04$ , BF = 5.89). Specifically, observers entrusted uncalculating non-helpers with only 21.73% ( $SD = 29.24$ ) of their endowment, and calculating non-helpers with 26.77% ( $SD = 31.86$ ) of their endowment. Helpers were sent more than twice as much: calculating helpers were entrusted with 61.20% ( $SD = 32.02$ ) and uncalculating helpers received the most with 64.14% ( $SD = 32.72$ ). Hereby, the differences between uncalculating helpers versus uncalculating non-helpers ( $b = 42.41$ ,  $t(1060) = 15.53$ ,  $p < .0001$ ) and calculating helpers versus calculating non-helpers ( $b = 34.44$ ,  $t(1060) = 12.61$ ,  $p < .0001$ ) were statistically significant. Moreover, in Experiment 1 there was a main effect for helping ( $b = 34.59$ , 95% CI [30.87, 38.32],  $t(2608) = 18.43$ ,  $p < .0001$ ), as observers entrusted more than twice as much to helpers than to non-helpers, and in Experiment 5 observers sent significantly less to punishers than to non-punishers ( $b = -7.30$ , 95% CI [-11.09, -3.52],  $t(2608) = -3.81$ ,  $p = .0001$ ).

Furthermore, I found no evidence to suggest that deliberation had a larger effect on actual trustworthiness for punishers compared to non-punishers, as the interaction effects were non-significant in Experiment 2 (H9.2a;  $b = 1.67$ , 95% CI [-3.96, 7.29],  $t(1305) = 0.58$ ,  $p = .56$ , BF = 0.14), Experiment 3 (H9.2b;  $b = 1.85$ , 95% CI [-2.29, 6.00],  $t(1301) = 0.88$ ,  $p = .38$ , BF = 0.3) and Experiment 5 (H13.2;  $b = 1.42$ , 95% CI [-5.17, 8.01],  $t(1302) = 0.42$ ,  $p = .67$ , BF = 0.13). For participants with excellent comprehension there were, however, main effects for both punishing ( $b = 6.33$ ,  $t(615) = 2.13$ ,  $p = 0.03$ ) and checking ( $b = -4.49$ ,  $t(615) = -2.30$ ,  $p = 0.02$ ) in Experiment 2, with punishers and uncalculating decision makers returning significantly more than non-punishers and calculating decision makers.

In Experiment 1 (H9.1) the effect of uncalculating decision making on trustworthiness was significantly larger when Players A decided to help compared to when Players A decided not to help ( $b = -10.81$ , 95% CI [-17.16, -4.45],  $t(1307) = -3.34$ ,  $p < .001$ , BF = 18.29). In line with predictions, non-helpers who did not check the personal cost of helping were the least

trustworthy, returning only an average of 15.69% ( $SD = 22.88$ ), whilst cost-checking non-helpers returned 21.29% ( $SD = 22.95$ ). Cost-checking helpers were substantially more trustworthy, returning an average of 43.54% ( $SD = 19.34$ ), whilst helpers who did not check the personal cost returned the most, with an average of 48.74% ( $SD = 19.34$ ). Post hoc tests on the estimated marginal means, accounting for multiple comparisons with the multivariate t-test (mvt) adjustment, revealed significant differences between uncalculating helpers and uncalculating non-helpers ( $b = 33.05$ ,  $t(1307) = 12.40$ ,  $p < .0001$ ), calculating helpers and calculating non-helpers ( $b = 22.25$ ,  $t(1307) = 12.11$ ,  $p < .0001$ ) as well as calculating helpers and uncalculating helpers ( $b = -5.21$ ,  $t(1307) = -3.74$ ,  $p < .001$ ), but not for calculating non-helpers and uncalculating non-helpers ( $b = 5.60$ ,  $t(1307) = 1.92$ ,  $p = .18$ ).

Conversely, and against predictions, in Experiment 4 (H13.1) the effect of uncalculating decision making on trustworthiness was significantly *smaller* when Player A decided to help compared to when Player A decided not to help ( $b = 11.04$ , 95% CI [-18.10, -3.98],  $t(1307) = -3.07$ ,  $p = .002$ ,  $BF = 8.59$ ). Nevertheless, in line with predictions, non-helpers who did not check the impact of helping were the least trustworthy, returning only an average of 17.24% ( $SD = 26.56$ ) of the endowment observers entrusted them with, whilst calculating non-helpers returned an average of 25.44% ( $SD = 23.84$ ), calculating helpers an average of 45.40% ( $SD = 18.57$ ) and uncalculating helpers an average of 48.23% ( $SD = 19.04$ ). Hereby, the differences between uncalculating helpers and uncalculating non-helpers ( $b = 31.00$ ,  $t(1307) = 10.13$ ,  $p < .0001$ ), calculating helpers and calculating non-helpers ( $b = 19.96$ ,  $t(1307) = 10.54$ ,  $p < .0001$ ), as well as uncalculating non-helpers and calculating non-helpers ( $b = 8.21$ ,  $t(1307) = 2.51$ ,  $p = .04$ ) were statistically significant.

### **6.5.8 Additional Unregistered Analyses: Does deliberation have a differing influence on trust and trustworthiness for non-helpers versus non-punishers?**

To provide a comprehensive perspective, a final unregistered analysis tested whether the effect of uncalculating behaviour on trust and trustworthiness differs for non-punishers compared to non-helpers. For trust there was no interaction between deliberation and behaviour ( $b = 2.22$ , 95% CI [-0.64, 5.08],  $t(2608) = 0.80$ ,  $p = .42$ ,  $BF = 0.98$ ), nor a significant main effect for deliberation ( $b = -1.27$ , 95% CI [-3.39, 0.85],  $t(2608) = -0.65$ ,  $p = .52$ ). However, observers sent

significantly more of their endowment to non-punishers than to non-helpers ( $b = -22.77$ , 95% CI  $[-26.70, -18.85]$ ,  $t(2608) = -11.64$ ,  $p < .0001$ ).

Furthermore, non-punishers returned a significantly higher proportion in the Trust Game than non-helpers did ( $b = -26.82$ , 95% CI  $[-32.40, -21.24]$ ,  $t(1007) = -9.43$ ,  $p < .0001$ ), and non-actors who checked the cost returned significantly less than those who made an uncalculated decision not to help/punish ( $b = -4.00$ , 95% CI  $[-7.09, -0.91]$ ,  $t(1007) = -2.54$ ,  $p = .01$ ). There was also a significant interaction between deliberation and experiment ( $b = 9.60$ , 95% CI  $[2.53, 16.66]$ ,  $t(1007) = 2.67$ ,  $p = .008$ ,  $BF = 3.45$ ), with uncalculated non-helpers returning the least ( $M = 15.69\%$ ,  $SD = 22.88$ ), followed by calculated non-helpers ( $M = 21.29\%$ ,  $SD = 22.95$ ), calculated non-punishers ( $M = 38.51\%$ ,  $SD = 22.42$ ), and uncalculated non-punishers ( $M = 42.51\%$ ,  $SD = 21.65$ ). However, when only participants with excellent comprehension were included, there no longer was a significant interaction ( $b = 7.89$ , 95% CI  $[-2.40, 18.17]$ ,  $t(486) = 1.51$ ,  $p = .13$ ,  $BF = 0.46$ ) although the average percentages returned remained similar (uncalculated non-helpers: ( $M = 13.74\%$ ,  $SD = 25.00$ ), calculated non-helpers: ( $M = 17.14\%$ ,  $SD = 20.61$ ), calculated non-punishers: ( $M = 36.27\%$ ,  $SD = 21.50$ ), uncalculated non-punishers: ( $M = 40.76\%$ ,  $SD = 20.06$ ).

## 6.6 Discussion

Previous work (Jordan et al., 2016) has shown that helping behaviour that is performed in a reflexive or uncalculating manner can yield reputation benefits, since observers infer that these actions reflect genuinely prosocial motives, rather than stemming from rational calculation of costs and benefits. Accordingly, uncalculated help signals trustworthiness and people are more likely to behave in an uncalculated way when they are observed (Jordan et al., 2016). Over five experiments, I replicate this study and extend it by examining whether uncalculated punishment also leads to reputation improvements. In a further extension of previous work, I also ask whether punishers and helpers deliberate over the *impact to the target* (rather than the personal cost to themselves) and how such ‘impact deliberation’ is viewed by bystanders. In Study 1 (personal cost deliberation) I expected both uncalculated help and punishment to signal trustworthiness. In Study 2 (target impact deliberation) I expected uncalculated help to signal trustworthiness. Conversely, I expected *calculated* punishment to signal trustworthiness. As punishment inflicts harm on another, I expected that people would observe an implicit moral directive to deliberate over the harm they could inflict on another individual – and that individuals who inflict harm reflexively would be viewed negatively. Replicating previous results (Jordan et al., 2016), I found that uncalculated help signals trustworthiness: helpers who did not consider the personal cost of helping were both more trusted and trustworthy than helpers who deliberated over the cost. Punishment results were more mixed. Although punishers were more likely to perform uncalculated actions when observed, uncalculated punishment was not reliably associated with either perceptions of trustworthiness or with trustworthiness itself. Only uncalculating *non*-punishers were more trustworthy than calculating non-punishers. In contrast to the cost checking context, considering the impact of helping had a larger impact on the trust and trustworthiness of non-helpers than helpers. Lastly, I found no conclusive evidence to suggest that checking the impact of punishing influences perceived or actual trustworthiness.

In Experiment 1, I replicated Jordan et al.’s (2016), finding that uncalculating helpers were perceived as significantly more trustworthy than calculating helpers. Uncalculated helping provides a reliable signal of trustworthiness as it indicates that people are not considering the personal costs of helping and that helping stems from other-regarding rather than strategic motives. As in Jordan et al (2016), people were sensitive to these reputation benefits and were less likely to check the personal cost of helping when their decision process was observed than when it was hidden (H1.1). Finally, as in Jordan et al. (2016), these reputation benefits were

restricted to those who helped: deliberation had no effect on trust (H7.1) or trustworthiness (H15.1) when participants decided *not* to help.

I similarly expected uncalculated punishers to be perceived as, and to actually be, more trustworthy than those who deliberated over the personal cost of punishing (Exp. 2 & 3). I also expected people to be sensitive to these reputation consequences and to be less likely to check the personal cost (or to decide more quickly) when observed. These predictions were only partially supported. Participants were half as likely to check the cost of punishing when their decision process was observed (H1.2a) and were also significantly faster in their decision-making (H1.2b). In contrast to predictions, observers directionally trusted *calculating* punishers more than uncalculating punishers (while the Bayes Factor value for H2.2a indicated support for this effect, results for H2.2a and H2.2b were statistically non-significant). Trustworthiness results were also mixed. Uncalculating punishers were directionally more trustworthy than calculating punishers in Experiment 2 (H14.2a), but directionally less trustworthy in Experiment 3 (H14.2b). Note that Bayes Factor values indicate support for the null hypotheses, although power requirements were not met for H14.2a and H14.2b.

Whilst I expected both uncalculated help and uncalculated punishment to signal trustworthiness, I had diverging predictions around decisions *not* to act. Decision conflict over whether to help/punish could stem from self-interested considerations of whether to pay a cost. But unlike helping, punishment decision conflict could also stem from concerns about inflicting harm on the target. As participants initially believed punishing could potentially be free, I expected uncalculating non-punishers to be perceived as harm averse. Conversely, calculating decisions not to punish would indicate a selfish decision (the personal cost of punishing being too high). Support for these predictions was mixed. As expected, uncalculating non-punishers were more trustworthy than calculating non-punishers (though effects were only significant for Exp. 2, H15.2a and not Exp. 3, H15.2b). Perceived trustworthiness was not reliably affected, as observers directionally trusted uncalculating non-punishers more in Experiment 2 (H7.2a; the Bayes Factor value indicates null findings are inconclusive) but directionally trusted calculating non-punishers more in Experiment 3 (H7.2b; the Bayes Factor value indicates support for the null hypothesis).

Uncalculated punishment does not therefore seem to be perceived as a signal of trustworthiness – and uncalculated punishers were not more trustworthy. As deliberative decisions are often considered to be wise (Bazerman, 2006; Grossmann et al., 2017; Pinker, 2022), uncalculated punishment might conceivably signal reduced competence (Jordan & Kteily, 2023), which

could have affected perceived trustworthiness. While possible, this explanation is unlikely as the same ought to have been true for the helping context in Experiment 1. Alternatively, it is possible that the signalling effect of uncalculating punishment was too small to have been captured by the present work. However, several of the Bayes Factor values for null results in Experiments 2 and 3 were less than 0.33, supporting the absence of an effect as opposed to a need for more data. Moreover, I frequently found the directional opposite of predictions, especially when deliberation was operationalised as decision time.

Helping may enhance reputation more than punishment because, even though third-party punishment is often viewed as a morally justified form of harm, people may still be unsure about those who engage in it (Dhaliwal et al., 2021; Raihani & Bhhary, 2015b, 2015c). Observers may therefore be unsure whether to trust punishers over non-punishers in the first place. Prior research has found that non-punishers can sometimes be trusted as much as punishers (Batistoni et al., 2022; Przepiorka & Liebe, 2016), and occasionally third-party punishers are even trusted less than non-punishers (Eriksson et al., 2016; Rai, 2022; Heffner & FeldmanHall, 2019; Strimling & Eriksson, 2014). I found no significant difference in the perceived trustworthiness of punishers and non-punishers. Nevertheless, trustworthiness did vary. When restricting the sample to participants with excellent comprehension, punishers returned significantly more than non-punishers.

Perhaps punishment needs to be seen as the ‘right thing to do’ for the decision process to matter as a signal. This can be difficult, as punishment – unlike helping – is morally bad when undeserved, and there are also questions around legitimacy in that a fellow participant in an economic game may not be seen as an appropriate person to intervene (Eriksson et al., 2017; Strimling & Eriksson, 2014). Further, it has been argued that defection in economic games can be considered ‘fair game’, making the punishment of it less justified (Dhaliwal et al., 2021; Salcedo García, 2020). This is additionally important because the appropriateness of non-action decreases for more serious infractions (Eriksson et al., 2021). Furthermore, even third-party punishers can be perceived as spiteful or competitive rather than prosocial, particularly when punishment is excessive (Dhaliwal et al., 2021). However, punishment in this study is unlikely to be seen as excessive: punishing still leaves the defector with £0.15, the amount that they would have received had they themselves acted fairly. It is also unlikely that punishers are perceived as being competitive (aiming to increase their own payoffs relative to the defector’s; Bone & Raihani, 2015; Price et al., 2002; Raihani & Bshary, 2019) because, whether participants choose to punish or not, they always end with only one-third of the amount the



defector receives (when participants punish, they finish the game with £0.05, whilst the defector finishes with £0.15, and when participants do not punish, they finish with £0.10, whilst the defector finishes with £0.30).

It should be noted, that although helpers were indeed trusted significantly more than punishers (H4), they were not actually more trustworthy than punishers (H18). The additional unregistered analyses further showed that non-punishers were significantly more trustworthy than non-helpers, which is in line with prior research (Batistoni et al., 2022). Just as help is a more reliable signal of trustworthiness than punishment, not helping is a more reliable signal of *untrustworthiness* (Raihani & Bshary, 2015b).

I expected deliberating about the *impact* of help / punishment to reflect negatively on helpers but positively on punishers, on the basis that considering both the prosocial aspect of third-party punishment and the negative consequences to the defector may be perceived as the ‘right’ thing to do when making punishing decisions. Helping on the other hand should be perceived as the ‘right’ thing to do, whether it helps a little or a lot. Helping without checking the impact to targets was therefore expected to signal trustworthiness. Conversely, deciding not to help without this consideration was predicted to result in particularly negative evaluations.

As predicted, uncalculating helpers were both more trusted (H6.1) and trustworthy (H19.1) than calculating helpers – but only directionally and with inconclusive Bayes Factors. Further, non-helpers who made calculating decisions were indeed more trustworthy than uncalculating non-helpers, but only significantly so when all participants were included (note: power requirements were not met for H20.1). Non-helpers were also directionally more trusted when they made calculating compared to uncalculating decisions (H11.1; the effect was statistically non-significant but supported by the Bayes Factor value). Counter to predictions, impact checking had a larger effect on the trust (H12.1) and trustworthiness (H13.1) of non-helpers compared to helpers. I did not find this result to be especially surprising. Indeed, I also argued that helping, regardless of the impact, is what matters the most for perceptions of (and actual) trustworthiness. It is possible that impact consideration matters more for those who decided *not* to help, as those who consider the impact at least considered helping, whereas those who do not consider the impact and do not help may be perceived as unwilling to help no matter how much another may benefit from it.

As predicted, considerations of target impact generally produced the opposite pattern in the punishing context. Counter to predictions, participants with excellent comprehension were significantly *less* likely to check the impact of punishing when they were observed than when

their decision process was hidden (H5.2). Directionally, punishers who checked the impact to the target were perceived as (H6.2) and directionally were (H19.2) more trustworthy than uncalculating punishers (though effects were statistically non-significant). No conclusive results were found for non-punishers (H11.2 & H20.2). Bayes Factor values of non-significant analyses mostly indicated support for the null hypotheses (except H11.2 and H12.2, which were inconclusive, and H6.2 which indicates support for the alternative hypothesis).

The findings presented here should be interpreted within the context of certain limitations, particularly regarding ecological validity. The experimental design was highly abstract and therefore may not have fully captured the complexities of real-world decision-making processes related to trust and trustworthiness. Future studies could enhance ecological validity by employing scenarios or tasks more directly applicable to everyday situations, thereby potentially yielding more representative results. Additionally, past research emphasizes the importance of motive attributions in shaping evaluations of helpers and punishers (Jordan et al., 2016; Raihani & Bshary, 2015b; Raihani & Power, 2021), which the present study did not explicitly explore. Understanding the motives observers attribute to actors, as well as eliciting self-reported motives behind actors' calculated versus uncalculated decisions, could provide further insight into the mechanisms underlying third-party punishment and the extent to which it can be interpreted as a prosocial act. This could also help to differentiate between punishment and helping as signals of trustworthiness.

Lastly, there may be some concerns inherent to the Trust Game itself. I used this game to measure attitudes towards punishers as previous work has shown that punishment increases trustworthiness whereas results on whether punishers are 'liked' or rewarded for their actions are more mixed (Barclay, 2006; Horita, 2010; Kiyonar & Barclay, 2008; Raihani & Bshary, 2015c). Nevertheless, it is important to acknowledge that the Trust Game does not fully disambiguate between trust and other underlying mechanisms. While decisions in the Trust Game can be partly attributable to risk attitudes (Chetty et al., 2021), decision patterns in Trust and Risk games differ (Houser et al., 2010). Differences in responses to help and punishment in the present study also indicate that observers did not make decisions based purely on risk preferences (if risk preferences were the key driver behind Trust Game decisions, I would not expect to observe any differences in help / punishment conditions).

In summary, the present chapter found that acting without considering personal costs signals trustworthiness in helpers but not punishers, and that considering the impact of help and punishment likely has little or no influence on perceived or actual trustworthiness.

# Chapter 7

## General Discussion

## 7.1 Summary of Context and Rationale

Prosocial behaviours encompass a wide range of actions (e.g., helping, volunteering, and certain punitive behaviours) that benefit others but often come at a personal cost to the individual performing them (Penner et al., 2005). These behaviours are crucial for the effective functioning of interpersonal relationships, groups, and societies. On a global level, prosocial actions play a significant role in addressing major challenges such as climate change mitigation (Alston, 2015; Marx & Weber, 2012) and pandemic management (Góis et al., 2019; Jin et al., 2021). As these behaviours involve personal sacrifices, they present a fundamental puzzle for evolutionary theory. Recognising and empirically demonstrating that humans can solve this puzzle through the mechanism of reputation has been a significant scientific achievement (e.g., Alexander, 1979, 1987; Barclay, 2012; Nowak & Sigmund, 1998, 2005; Suzuki & Akiyama, 2005; Wu et al., 2016a, b). Indeed, attributing reputations to others, sharing reputational information, and basing behaviour on these reputations are key drivers of large-scale human cooperation (Giardini & Wittek, 2019; Milinski et al., 2002; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2004). However, most psychological research on reputation stems from artificial, abstract laboratory settings, such as economic games. While these settings are well-controlled, they lack many aspects of the real-world environments where human behaviour naturally occurs. This thesis therefore conducted 10 studies involving more than 14,700 participants, with a particular focus on cooperative and punitive behaviours, using several methods with higher ecological validity, to study reputation in everyday life.

## 7.2 Summary of Findings

Prior research on reputation management has typically focused on single reputation domains or behaviours in isolation, often conducted in highly controlled and nearly context-free settings. **Chapter 2** reports that, using a daily diary phone app, participants felt significantly more concerned about their reputation during situations where they were managing their reputation compared to other social interactions. In line with previous research on individual factors, reputation concern was influenced by (i) the social connectedness of others, (ii) the actor's relationship to the other person(s), (iii) the value placed on the person(s) whose view they were managing, (iv) the closeness felt towards the person(s) whose view they were managing, and (v) the likelihood of future interactions. A more holistic analysis method (fitting a singular model including all predictors of reputation concern rather than multiple individual models)

showed that social relationship factors (i.e., relationship value and relationship type) influenced reputational concern more than other cues.

**Chapter 3** used a weekly survey study conducted across six weeks to explore how punishers' reputations vary in everyday life and to identify key factors affecting these reputations. Except for the type of punishment (third-party punishment (TPP), second-party punishment (SPP), or a combination thereof), every context factor influenced the reputation of punishers (at least directionally) as predicted. Even for the 'type of punishment' factor, I argued that the premise holds: punishers acting out of concern for others are viewed more positively than those with self-serving motives. Reputations were mainly influenced by attributed motives and perceived justification of the punishment, with the punisher-violator relationship also playing a significant role. The severity of the violation was the weakest predictor, affecting only perceptions of a punisher's toughness. Punishment seemed to offer distinct reputational benefits in terms of prosociality and formidability, especially considering the punisher-violator relationship. These findings highlight the importance of contextual factors in understanding the reputational consequences of punishment.

**Chapter 4** examined individuals' awareness of how their cooperative and uncooperative behaviours are perceived by others, along with the accuracy of observers' evaluations. I found that actors accurately anticipated how their cooperative and uncooperative actions would be perceived, although they believed that their motives were more apparent to others than they actually were. Actors also systematically underestimated how positively their uncooperative behaviours would be judged. Both observers and actors exhibited higher normative accuracy for cooperative compared to uncooperative behaviours, but displayed higher distinctive accuracy (i.e., perceptions of the *unique* traits and motives relative to other actors) for uncooperative behaviours. Contrary to my predictions, cooperative behaviours were seen as more indicative of character than uncooperative behaviours. This chapter highlights the complex dynamics of behaviour perception in social interactions, revealing nuanced patterns in perceptive accuracy and biases.

**Chapter 5** used a vignette study to explore the evaluation of, and motives attributed to, peer punishers in third-party versus second-party contexts. As predicted, third-party punishers were viewed more favourably than second-party punishers, with observers attributing prosocial motives to third-party punishers and self-centred or formidability concerns to second-party punishers. Participants believed that SPP and TPP were highly driven by deterrence and communication of disapproval, and rejected the idea that punishers sought reputation benefits.

The second study in Chapter 5 replicated these findings with a more complex method and additionally showed that punishers anticipated these observer attributions, thereby demonstrating meta-perceptive accuracy. Similarly, observers accurately assessed punishers' motives and traits. Contrary to predictions, actors systematically overestimated the positive influence of third-party punishing on their prosocial reputation and underestimated its influence on their perceived formidability. There was no systematic bias in second-party punishing situations. Conversely, observers exhibited a negativity bias in both contexts, perceiving punishers as more motivated by formidability or self-centred concerns than they actually were. Political orientation also influenced punishing sentiments, with more conservative individuals viewing punishers more negatively and punishment as less justified. Findings from this chapter provide insights into how individuals perceive and interpret punishing actions, and the extent to which actors understand this.

**Chapter 6** explored the reputation benefits of uncalculated helping and punishment, extending previous research on helping behaviours. Through five experiments, the chapter examined whether uncalculated punishment, like uncalculated helping, signals trustworthiness, and investigated the effects of deliberation over both personal cost and impact on the target. The findings showed that uncalculated helping consistently signals trustworthiness, as helpers who did not consider personal costs were more trusted and trustworthy. Actors were sensitive to these reputation benefits, with participants less likely to check the personal cost of helping when observed. Although uncalculated punishment did not reliably signal trustworthiness, non-punishers who did not consider costs were more trustworthy. In contrast to the cost checking context, considering the impact of helping had a larger impact on the trust and trustworthiness of non-helpers than helpers. There was no conclusive evidence to suggest that considering the impact of punishing influenced perceived or actual trustworthiness. This chapter highlights the complexities in how different prosocial and punitive behaviours influence reputational outcomes.

## **7.3 General Implications**

### **7.3.1 Advancements in Research and Theory**

One of the primary contributions of this thesis is its emphasis on ecological validity in psychological research on reputation. Ecological validity—the extent to which the findings of a research study can be generalised to real-world settings—is particularly crucial for

psychology experiments aiming to bridge the gap between controlled environments and everyday life contexts. Researchers have additionally criticised the tendency to view reputation as a uni-dimensional variable, which often focuses on financial consequences only (Chen & Xu, 2020). Furthermore, researchers have highlighted that reputation is not constructed in a social vacuum and that we lack an understanding of how reputation is determined from the ‘muddle of imperfect, biased inputs’ people receive (Giardini et al., 2022). By employing methods such as daily diaries, weekly surveys, and vignette experiments, this thesis captured behaviours in more naturalistic settings, thereby yielding insights that are more applicable to real-world scenarios.

Importantly, the studies in this thesis reveal that reputation mechanisms observed in controlled laboratory settings are fairly comparable to those in more complex, noisy and naturalistic environments. Furthermore, my research underscores the significance of context and the necessity of simultaneously considering multiple situation, motivation, relationship, and other context factors influencing reputation and reputation management, rather than examining these factors in isolation. While many findings from controlled settings are corroborated in real-world scenarios, various nuances, deviations, and extensions emerge. This demonstrates that while controlled experiments are crucial for establishing foundational principles, a comprehensive understanding of reputation dynamics requires consideration of the variability and complexity inherent in everyday interactions.

Chapters 2 and 3, in particular, demonstrated the influence of factors previously tested in controlled environments when examined within the noisier contexts of everyday life. For instance, Chapter 2 showed that even without restrictions on the reputation management behaviours, relationships, or reputation domains sampled, findings largely aligned with previous literature. When contextual factors were considered individually, all but one predictor variable (audience presence) influenced reputation concern in a manner consistent with laboratory or field findings. The exception of audience presence was likely due to participants focusing on whether others witnessed their behaviour, rather than the distinction between their actions becoming public knowledge versus remaining private, as commonly emphasised in economic game studies. This highlights the importance of being clear about such nuances or distinctions. More importantly, Chapter 2 underscored the value of examining contextual factors together, rather than in isolation. By doing so, it was possible to determine the relative importance of context factors and reveal that only two context factors (relationship value and relationship type) remained as statistically significant influencers on reputation concern when

examined collectively. Chapter 2 therefore extended previous findings by demonstrating that aspects of social relationships influence reputation concern more than social visibility cues.

Chapter 3 showed that even with a wide range of known and unknown interaction partners from a multitude of real-life situations, and without the predetermined options of the laboratory, findings about the factors influencing the reputation of punishers mostly supported previous research. Furthermore, examining a broader context not only enabled me to confirm prior findings on the influence of the punisher-violator relationship on a punisher's reputation but also to extend them. Previous research indicates that punishers are judged more favourably when they are socially closer to the violator, while those who punish strangers receive more negative evaluations (Eriksson et al., 2017; Strimling & Eriksson, 2014). My research clarified this by showing that it is a punisher's *prosocial* reputation which can improve when close others are punished. However, rather than gaining prosocial reputation benefits, punishers are more likely to be perceived as stronger and more dominant for punishing, and especially so when punishing individuals less close to them. It is primarily the formidable aspect of their reputation that is enhanced, with prosocial reputation benefits occurring under specific circumstances. This aligns with predictions from Raihani and Bshary (2015b) that punishment can affect reputation by signalling that a punisher is competitive, or under restricted circumstances, has cooperative tendencies or fairness preferences. Moreover, Chapter 3 found that, in line with previous literature, power dynamics primarily influence perceptions of formidability (Gordon & Lea, 2016). It also clarified that the impact on prosocial reputation aspects is weak and mixed. This further emphasises that research should distinguish between different reputation domains and include context-specific factors for a fuller understanding of reputation benefits, as there may be distinct strands of reputational advantages.

This thesis also focused on comparing second- and third-party punishment and their influence on reputation. According to previous research, third-party punishers should be more likely to receive more prosocial reputation benefits than second-party punishers, as the former generally do not stand to directly benefit from punishing and are thereby viewed as other-regarding whilst second-party punishers are often thought to be acting out of self-interest (e.g., Fehr & Gächter 2002; Martin et al., 2019; Raihani & Bshary, 2015b; Raihani & McAuliffe 2012). Consequently, TPP is frequently associated with cooperative intentions (e.g. Jordan et al., 2016; Jordan & Rand, 2017; Raihani & Bshary, 2015b,c), whereas SPP is linked to competitive and self-centred motives (Carpenter, 2003; Peysakhov et al., 2014; Raihani & Bshary, 2015b; Yamagishi et al. 2012). The first study in Chapter 5, which used 14 vignettes, supported these



findings: TPP was regarded more favourably than equivalent SPP actions. The second study, which solicited nearly 300 descriptions of everyday punitive actions participants recently engaged in, confirmed this. Reputation benefits were stronger for TPP than SPP and were not limited to prosocial traits (e.g., moral, cooperative, trustworthy) but also included traits such as strength, competence, and dominance. Both studies found that observers associated TPP with prosocial motivations, whereas SPP was associated with self-protection and formidability concerns. However, it is important to note that the findings from this chapter may not necessarily apply to targets of punishment (or victims in TPP scenarios), as participants were neutral third parties. Violators and victims may evaluate punishers differently. The study in Chapter 3 (in which participants could be an observer, victim, or violator in punitive instances) further highlighted the importance of considering other context factors. Contrary to previous literature (e.g., Barclay, 2006; Jordan et al., 2016; Jordan & Rand, 2020; Kurzban et al., 2007; Nelissen, 2008; Raihani & Bshary, 2015b, 2015c), the study in Chapter 3 found no evidence to suggest that third-party punishers are evaluated more positively than second-party punishers when other contextual factors (e.g., perceived justness and motives of punishment, violation intentionality and severity, power dynamics, and the violator-punisher relationship) are also considered. However, the underlying foundation of prior predictions (Raihani & Bshary, 2015b) was supported: the attribution of selfish or altruistic motives was key to reputation consequences. Although third-party punishers were viewed as acting more altruistically than second-party punishers, in a noisy environment with multiple factors, the type of punishment alone was not a strong predictor of how a punisher would be evaluated. Previous studies corroborate this, finding that even third-party punishers can be seen as less interested in fairness restoration and more in harming a violator when factors such as the proportionality of the response are considered (Dhaliwal et al., 2021). Therefore, while it is valuable to differentiate between SPP and TPP, and third-party punishers do tend to be attributed more altruistic motives than second-party punishers, the variability within specific SPP and TPP contexts may influence reputation more than whether the punisher was affected by the initial transgression or not. Focusing on the distinction between SPP and TPP alone can therefore be too simplistic when considering the complexity of real-world punishing incidents. For instance, a third-party punisher who punishes a competitor for a minor infraction would likely be evaluated less positively than a second-party punisher who justly takes someone to court for a major violation. Still, the same key conclusion remains: reputation consequences depend on the information observers can extract regarding the potential motives underpinning the decision to punish (Brañas-Garza et al., 2014; Nelissen, 2008; Raihani & Bshary, 2015b).

Chapter 6 additionally indicated that the inference of selfish motives may be more important than other motive cues. Past research has highlighted the importance of motive attributions in shaping evaluations of both helpers and punishers (Jordan et al., 2016; Raihani & Bshary, 2015b; Raihani & Power, 2021), with people giving positive recognition based on the perceived purity of an actor's motives (Berman & Silver, 2022; Raihani & Power, 2021). Chapter 6 replicated the finding that observers might suspect that helping stems from strategic rather than altruistic motives when the helper deliberates over personal costs before deciding whether to help (Hoffman et al., 2015; Jordan et al., 2016). I further found that this effect of deliberation does not extend to costly punishment. While personal cost deliberation may signal trustworthiness in those refraining from punishment—perhaps because it shows an unwillingness to harm another out of principle—when directly tested, deliberation about the impact on the target of help or punishment had no effect on trustworthiness. Therefore, cues that reduce the likelihood of selfish motives may be more important than cues of regard for the target of the behaviour. Furthermore, previous research has argued that helping may enhance reputation more than punishment because, even though third-party punishment is often viewed as a morally justified form of harm, people may still be unsure about those who engage in it (Dhaliwal et al., 2021; Raihani & Bshary, 2015b,c). The complexity observed in uncalculated punishment in Chapter 6 confirms a more nuanced role of punitive behaviours in reputation, expanding current psychological models to account for yet another context-dependent variation. Even specific context cues—such as calculated versus uncalculated behaviours—are not "one size fits all", as their effect on trustworthiness varied across behaviours and the aspect (personal cost versus target impact) being deliberated about.

Another identified research gap (Raihani & Power, 2021) that this thesis touched upon is the cognitive abilities underlying reputational strategies. Both reputation management and impression formation depend on a range of cognitive skills. To form accurate impressions of others, we must be capable of inferring their motives (here demonstrated via observer judgement accuracy). Likewise, when managing our reputation, it is crucial to accurately predict how our behaviour, along with various interpretations of our motives, will shape others' opinions of us (here demonstrated via actor meta-accuracy). Although we do not yet know which cognitive mechanisms are essential to these skills, the ability to manage one's own reputation and assess the reputations of others likely hinges on epistemic vigilance (Sperber et al., 2010). A crucial aspect of epistemic vigilance is the ability to discern intentions that contradict observed actions (such as when someone helps to gain reputation benefits rather than to be kind). This form of intention attribution is complex and is not observed in non-human

apes or very young children (Tomasello et al., 2005). The development of scepticism involves an even higher level of sophistication: we not only recognise that an individual's intentions might conflict with their actions but also suspect that the individual is attempting to deceive others by projecting a false motive (e.g., Alex wants me to believe they helped to be nice, but I suspect they actually want something in return) (Giardini et al., 2022). In Chapters 4 and 5, I explored both observer judgment accuracy and actor meta-accuracy, while also investigating whether perception accuracy varies across different behaviours.

Actors demonstrated meta-accuracy when predicting evaluations for their cooperative, uncooperative, second-party punishing, and third-party punishing behaviours. They understood how they would be uniquely perceived relative to the average actor and showed insight into potential misperceptions by observers (insight; Carlson, 2016), accurately predicting differences between their self-perceptions and observer judgements. For all behaviours examined, actors overestimated how accurately observers would perceive them (transparency bias; Gilovich et al., 1998). Consistent with error management theory (in which individuals favour strategies which minimise the more expensive error type by increasing the occurrence of the cheaper type; Efferson et al., 2020; Kiyonari et al., 2000; McKay & Efferson, 2010; Yamagishi et al., 2007), actors exhibited a systematic negativity bias when predicting evaluations based on their uncooperative behaviours, expecting observers to view them less favourably than they actually did. In contrast, third-party punishers displayed a systematic positivity bias, overestimating the positive impact of their punitive actions on their prosocial reputation and underestimating their influence on their formidability reputation. However, it is important to note that this positivity bias should be interpreted in context: third-party punishers were aware that observer perceptions would be more negative than their self-perceptions, but actual perceptions were even more negative than anticipated. The positivity bias should therefore not be viewed as a mere positivity bias in self-evaluation, as third-party punishers did have knowledge of the directional difference in observer versus self-perceptions. I found no systematic bias for cooperative or second-party punishing behaviours. These findings suggest that meta-accuracy may vary depending on the behaviour in question and that biases can persist even when individuals accurately predict perceptions. Systematic biases also appear to be context dependent. Overall, these findings contribute to understanding how individuals perceive and predict others' perceptions of their behaviours, shedding light on the complexity of reputational strategies and cognitive capabilities involved in managing one's reputation.

Observers, in turn, were adept at accurately assessing the actual motives and trait self-perceptions provided by actors. They consistently demonstrated insight across all four behaviours examined, accurately evaluating actors' true motives and traits beyond actors' expectations. However, observers also sometimes perceived certain items inaccurately and in ways that actors had anticipated. Systematic biases in observer perceptions were only explored in Chapter 5, which revealed a negativity bias in TPP and SPP contexts. Observers tended to believe that punishers were more motivated by formidability or self-centred concerns than they actually were, while underestimating their other-regarding motives. These findings align with prior research emphasising that punishment often benefits the punisher's situation over the violator's, leading observers to question the moral legitimacy of punishment (Fehr & Rockenback, 2003; Raihani & Bshary, 2015b; Xiao, 2013). This underscores the presence of epistemic vigilance, particularly the scepticism among observers regarding actor motives. Overall, the findings from Chapters 4 and 5 suggest that observers possess the ability to effectively engage in perspective-taking with actors, while actors can anticipate how they would be perceived by observers.

### **7.3.2 Practical Implications and Applications**

Reputations are a cornerstone of social life. Therefore, understanding how people infer and manage reputations in naturalistic settings can have practical implications for various stakeholders, including individuals, organisations, and policymakers. Designing and testing interventions based on the findings presented in this thesis could enable stakeholders to make more informed decisions. For example, recognising that individuals care more about their reputation when they value the other person can inform strategies to improve workplace dynamics. If organisations focus on fostering positive relationships among employees and between employees and leaders, individuals may be more motivated to maintain a good reputation. This could lead to increased cooperation, productivity, and overall workplace harmony (Suknunan & Bhana, 2022). Similarly, promoting or providing opportunities for uncalculated helping behaviours within teams might foster a more trusting and cooperative environment, which in turn may enhance productivity or job satisfaction (e.g., Brown et al., 2015; Hauser et al., 2016; Tjosvold, 1982). However, a key finding of this thesis is the context-dependent nature of reputation which warns against the assumption that strategies effective in one context will necessarily be successful in another. Stakeholders should test and adapt interventions to ensure their applicability and effectiveness in various settings.

Negatively evaluating prosocial behaviours can hinder large-scale cooperation by disincentivising cooperative behaviour (Raihani & Power, 2021). This thesis identified several factors that influence the reputations of actors. Being aware of these factors may be able to help people feel more secure in certain actions being socially rewarded. Conversely, making people aware of the biases and suspicions we possess against others' actions may help reduce stigma against well-intentioned behaviours. Simply acknowledging biases could potentially lessen their impact (Dunne et al., 2018; Liu et al., 2024; Pope et al., 2018; Sukhera et al., 2020; Zhu et al., 2022), though this idea requires further investigation in this context.

The complexity of punitive actions and resulting perceptions are also highlighted in my findings. These insights may help in designing policies and interventions involving punishment. As punishments perceived as unjust increase employee misconduct (Mooijman & Graham, 2018), understanding how the reputational consequences of punishment are influenced by who administers it and under what circumstances may be useful. Considering factors such as attributed motives and the punisher-violator relationship could help develop disciplinary measures that have fewer negative repercussions.

This thesis further highlights the importance of motive attributions behind behaviours and their impact on reputational outcomes, which can have implications for leadership and management practices. Indeed, leaders' reputations are influenced by the perceived justness and motives behind their decisions (Hall et al., 2004). Recognising observer biases in interpreting motives from leaders and coworkers may lead to better communication strategies to manage perceptions. Additionally, understanding that actors' motives may not be as transparent to observers as assumed can encourage more effective communication, which may in turn foster smoother relationships, trust, and productivity (Bharucha, 2018; Boies et al., 2015; Kalogiannidis, 2020; Kim et al., 2017; Odine, 2015; Sadia et al., 2016).

Finally, the insights from this thesis can be integrated into educational courses in psychology or behavioural economics. Incorporating real-world context studies into these programmes can offer a more comprehensive understanding of human behaviour and equip students with broader methodologies. Such courses could highlight the distinctions and connections between controlled experimental studies and real-world research, better preparing students for the practical application of these methods.

## **7.4 General Limitations**

The limitations of the individual studies within this thesis have been addressed in the discussion sections of previous chapters. In this section, I address the broader limitations of the overall approach taken in this thesis.

#### **7.4.1 Experience Sampling Methodologies**

Firstly, while the experience sampling methodologies used in this thesis are valuable for documenting social behaviours and the contexts and relationships in which they occur, these methods rely on self-reports. Self-reporting can introduce biases, such as social desirability bias or inaccurate self-perception. For instance, despite responses being anonymous in all studies, participants in Chapters 2-5 may have been influenced by (conscious or subconscious) reputation management motives, leading to selective reporting. Participants might have refrained from describing situations they perceived as embarrassing or morally bad to avoid negative evaluations, instead opting to highlight instances that cast them in a more favourable light. Conversely, it is possible participants might have been inclined to report more exciting or noteworthy events to avoid appearing uninteresting, which could skew the data toward ‘larger’ incidents. They may also have disproportionately recalled and reported more salient events because these are typically more noticeable and memorable. (It should, however, be noted that many reported incidents did include mundane, embarrassing, and even morally questionable situations, although the frequency of these occurrences was not tested). Participants may also have exhibited biases in how they reported their actions. For example, participants may have been more inclined to include mitigating explanations for uncooperative actions, emphasise personal sacrifice for cooperative ones, provide justifications to show that punishment was deserved in SPP cases, or actively portray actions in a more altruistic light for TPP. Many participants included explanations for their actions, possibly to reduce potential negative judgments or enhance positive ones, depending on the behaviour type. This may have introduced an additional bias, especially relevant in Chapters 4 and 5, as perceptions and their accuracy could be influenced by the framing and extent of the explanations provided. Furthermore, self-reports are inherently limited by participants' awareness of their motivations and behaviours. A lack of self-awareness can introduce bias, particularly if participants misinterpret or rationalise their behaviour retrospectively, resulting in discrepancies between reported and actual behaviours, motivations, or feelings.

Additionally, self-selection biases and attrition, common issues in most studies, may be exacerbated in experience sampling methodologies (Scollon et al., 2003), particularly in those described in Chapters 2 (involving daily sampling over one week) and 3 (involving weekly sampling over six weeks). The longer duration and increased effort required for these studies might result in certain types of individuals being overrepresented or underrepresented. Some individuals may refuse to participate outright, while less motivated participants might drop out quickly or lose interest once the novelty wears off. Consequently, the remaining participants may exhibit higher levels of motivation, conscientiousness, agreeableness, or other traits, leading to a non-representative sample (Scollon et al., 2003). After all, participants who ignore app notifications or forget to log back into Prolific every week are likely different from those who participate diligently. As a result, the data collected here may predominantly reflect the latter group.

Furthermore, most findings in this thesis do not establish causality. While experience sampling methodologies are useful tools for documenting social behaviours and their contexts, they are less effective in determining causal relationships.

A similar important consideration is the methodological constraints associated with balancing experimental control and ecological validity. While these studies aimed to enhance ecological validity by capturing behaviours and experiences in naturalistic settings, some degree of experimental control was therefore sacrificed. This trade-off may introduce potential confounding factors or limit the generalisability of certain findings to more controlled settings. Combining experience sampling methodologies with other methods, such as economic games, that can establish causal links and have high internal validity (Bardsley et al., 2010; Guala & Mittone, 2010; Loewenstein, 1999), therefore remains valuable.

Moreover, I should acknowledge that this thesis lacks the ecological validity found in many anthropological studies, which often include methods such as direct observations and network analyses. Various important aspects that would enhance the ecological validity of studying reputation and reputation management were not assessed in this thesis, including socio-ecological factors such as cultural influences and network fluidity. Combining experience sampling methodologies with other field methods, such as direct behavioural observations, ethnographic methods, and lab-in-the-field studies, could provide a more comprehensive understanding of reputation (Balliet et al., 2022).

## 7.4.2 WEIRD Participants and Socio-Ecological Factors

All of my studies were conducted online, predominantly using participants from WEIRD (Western, Educated, Industrialized, Rich and Democratic; Henrich et al., 2010) societies. This reliance on WEIRD samples may have affected the generalisability of my findings. Reputations are inherently shaped by the unique standards of each social network, dictating what constitutes "positive" or "negative" behaviours, traits, and intentions. As a result, comparing reputations across different networks is important, as what earns esteem in one context might be perceived negatively in another. Due to this contextual relativity, there exists no set of behaviours that universally guarantee a positive or negative reputation (Zinko et al., 2017). Instead, reputations are largely shaped by how closely one's actions align with or deviate from the norms of a specific social group (Zinko et al., 2017). Several theoretical perspectives highlight the influence of socio-ecological factors, such as pathogen prevalence (Enke, 2019; Fincher et al., 2008), population size (Marlowe et al., 2008, 2011), and ecological demands (Gelfand et al., 2011; Van De Vliert & Van Lange, 2019), in shaping norms and institutions that govern social behaviour. Other theories emphasise the role of cultural factors that influence the types and frequency of punishment across societies (Brauer & Chaurand, 2010; Eriksson et al., 2017; Schulz et al., 2019; Uskul et al., 2023). For instance, studies conducted in non-industrial, small-scale societies typically reveal that individuals in these communities are less inclined to engage in punitive actions, even in economic games (Henrich et al., 2010; Wiessner, 2009). Investment in third-party punitive actions is particularly rare (Marlowe, 2009; Marlowe et al., 2008, 2011). In these contexts, individuals who administer punishment may consequently receive fewer reputational rewards.

Additionally, Dumas et al. (2021) demonstrate the importance of broad samples by exploring how social capital or prominence can influence signalling dynamics. Incorporating such insights into analytical and agent-based models reveals 'reputational shields' and 'reputational poverty traps'. The 'reputational shield' scenario involves low-quality individuals being erroneously perceived as high quality due to their social standing, while the 'reputational poverty trap' suggests that high-quality individuals struggle to enhance their reputation due to a lack of social capital. Based on a case study focused on religious signalling in South India, there is a compelling argument for increased scrutiny of the complex interplay among reputation, prominence, and social capital, as well as the resultant impacts (Dumas et al., 2021).

Research indicating that preferences for social partners can be influenced by culture also underscores the need for caution when interpreting studies citing preference for cooperative



individuals as evidence for partner choice driving the evolution of cooperation (Smith & Apicella, 2020). Such studies often draw from Western samples, where conditions for partner choice may be particularly favourable. A study investigating qualities associated with being a preferred campmate among Hadza hunter-gatherers revealed that preferences changed over time. Initially, individuals preferred better hunters, but over time they prioritised more cooperative campmates (Smith & Apicella, 2020). Additionally, anecdotal evidence suggests that Hadza individuals with greater exposure to outside cultural institutions (e.g., schooling, employment, or village living) had a stronger preference for generous campmates (Smith & Apicella, 2020). This variability underscores the importance of recognising that the traits which confer reputational benefits can vary across different cultural contexts.

These studies underscore the limitation of not incorporating various socio-ecological factors and relying solely on WEIRD samples to understand the breadth of social behaviours and norms across diverse populations, which is essential for studying reputation dynamics. It is imperative to incorporate insights from cross-cultural research to discern both universal and culturally specific factors influencing reputation and reputation management.

### **7.4.3 Limited Investigations of Indirect Reciprocity and Reputation-Based Partner Choice**

Another limitation of this thesis is that it did not examine the (actual) subsequent behaviours of observers or the resulting fitness outcomes of the individuals involved. While some indirect measures of indirect reciprocity (IR) and reputation-based partner choice (RBPC) are incorporated in Chapters 3 and 6, this thesis did not explicitly investigate these mechanisms.

A person's reputation acts as an evaluation of their past behaviour, influencing how others interact with them in future encounters. As discussed in the general introduction, having a positive reputation can provide various advantages through two main pathways: individuals are more likely to be selected as partners (RBPC, Roberts, 1998) or they may receive rewards from others (IR, Kandori, 1992; Ohtsuki & Iwasa, 2007; Roberts et al., 2021). Although many models focus on a simplistic 'good' versus 'bad' reputation framework, real-world reputation systems are considerably more intricate (Barclay & Barker, 2020; Bartley & Child, 2014; Dafoe et al., 2014; Dukerich & Carter, 2000; Fombrun & Rindova, 1996; Heugens et al., 2004; Macfarlan et al., 2013; Rhee & Haunschild, 2006; Rogers et al., 2018; Tadelis, 2016).

Instead of directly observing instances of IR or RBPC in real-world contexts, my research focused more on the preliminary step of understanding when and why individuals acquire a positive reputation. By examining this foundation, I hope that future studies can later determine when a positive reputation leads to rewards or being chosen as an interaction partner.

Therefore, my research centred on evaluating the factors contributing to positive reputations, which may precede analysing the consequences of these reputations. Still, it is essential to consider how these positive evaluations translate to real-world interactions and whether they influence partner selection and cooperation. While identifying individuals' concerns about their reputation and contexts in which they receive varying evaluations is important, it does not equate to showing that reputation influences behaviour in ways predicted by indirect reciprocity and partner choice theories. Future research should therefore directly investigate RBPC and IR in applied contexts to understand the broader implications of reputation in social dynamics.

#### **7.4.4 Other Limitations of Scope**

While reputations can develop across various domains based on collective information about individuals' behavioural inclinations (Giardini & Wittek, 2019), research has predominantly focused on a narrow range of reputation domains. This thesis, too, was largely confined to cooperative and punitive behaviours, emphasising prosocial and formidable reputations. Chapter 2 is an exception to this, as it investigated reputation management in any behaviour, domain, or situation.

Furthermore, while this thesis explored both second- and third-party punishment across various scenarios and relationship dynamics, it did not encompass centralised punishment institutions, or group/consensus punishment. Experimental research highlights the significance of consensus in legitimising punishment, with collective punishment decisions being more effective in promoting cooperation (Cardenas, 2000; Casari & Luini, 2009; Eriksson et al., 2017; Ertan et al., 2009; Hilbe et al., 2014; Pfattheicher et al., 2018; Shinada & Yamagishi, 2007; Villatoro et al., 2014; Zheng & Nie, 2013). Outsourcing punishment to central authorities (i.e., 'centralising' punishment) may also enhance its legitimacy, although evidence on its effectiveness compared to peer punishment is mixed (Andreoni & Gee, 2012; Baldassarri & Grossman, 2011; Carpenter et al., 2012; Grieco et al., 2017; Nosenzo & Sefton, 2014; O'Gorman et al., 2009; Traulsen et al., 2012). However, many arguments applicable to peer punishment often extend to centralised punishment (Raihani & Bshary, 2015b; Schoenmakers et al., 2014; Yamagishi, 1986). For an

overview of the many functions of punishment, including institutionalised punishment, refer to Raihani and Bshary (2019).

Similarly, regarding cooperative behaviours, studies suggest that reputation scores may lose their ability to foster cooperation when assigned to groups rather than individuals, or in group contexts instead of dyadic interactions (Duca & Nax, 2018; Nax et al., 2015). Finally, numerous other contextual factors that may influence reputation have not been explored here, such as the impact of wealth on perceived motives (Wang et al., 2024) or gender differences in reputational strategies (e.g., Jones & Linardi, 2012).

## **7.5 Future Studies**

Besides the general suggestions derived from the limitations explored (e.g., using non-WEIRD samples, combining with fieldwork, and explicitly searching for IR/RBPC links in real-world scenarios) or mentioned in previous chapters, the current research could be expanded in many ways. The following will outline a few of these potential expansions.

### **7.5.1 Metacognition and Reputation**

While Chapters 4 and 5 have provided insights (i) into peoples' capacity as observers to accurately infer others' motives and (ii) as actors to anticipate how their actions will be perceived, many questions remain unanswered about the cognitive mechanisms underlying reputation management and impression formation. A next step could be to investigate whether individuals vary in their ability to both manage their own reputation and form accurate impressions of others. For instance, a previous study on immoral actions found that the accuracy of individuals as observers was a significant moderator of their meta-accuracy as actors in a subsequent task (Lees et al., 2022). Moreover, conducting standardised cognitive tests to measure metacognition, such as false belief and perspective-taking tasks, could help determine whether individuals with higher metacognitive performance are also better at anticipating others' judgments of their actions.

Another line of inquiry could investigate the impact of self-presentation strategies on reputation, as well as perceptive accuracy across those strategies. People are aware that their behaviours may signal self-interested or strategic rather than altruistic intentions, and therefore

aim to prevent such interpretations by others. One of the ways in which individuals manage their reputation is by hiding or promoting their behaviour. If an individual engages in an undesirable behaviour which they think may reflect on them unfavourably - which could in turn bring negative consequences such as being punished (Falk et al., 2005; Fehr & Gächter, 2002, 2000; Johnson et al., 2009; Nikiforakis & Engelmann, 2011) or being avoided as a social interaction partner - they may attempt to protect their reputation by hiding said unfavourable behaviour so that others cannot judge them for it. Alternatively, if an individual engages in a desirable behaviour which they expect will reflect on them favourably, they may want to advertise this behaviour in an attempt to receive reputational benefits, such as being chosen as a social interaction partner (Barclay, 2016; Baumard et al., 2013; McNamara et al., 2008) or being more likely to receive help (Bliege Bird & Power, 2015; Gurven et al., 2000; Lyle & Smith, 2014; Macfarlan et al., 2012, 2013; Power & Ready, 2018). However, it should be noted that individuals who advertise their prosocial actions may also face reputation damage (Raihani & Power, 2021; Srna et al., 2017). Therefore, individuals sometimes hide their prosocial behaviours to avoid being seen as virtue-signalling (Levy, 2021), or try to increase reputational benefits even further by appearing modest (Quillien, 2019). Similarly, they can promote uncooperative behaviours to gain desirable evaluations, such as status, strength or dominance (Gordon et al., 2014).

By adapting methods from Chapters 4 and 5, future studies could measure participants' tendencies to hide and promote certain behaviours, and how these actions influence their reputation. Such studies could reveal where perceptions are most accurate and could either focus on the most direct combinations (hiding uncooperative behaviours and promoting cooperative behaviours) or attempt to deepen understanding on topics such as modesty.

### **7.5.2 Making Use of Other Games: Cyberball**

Integrating different methodologies in the study of reputation can enhance the robustness and applicability of findings. Cyberball, a virtual ball-tossing game that simulates social interactions and induces real emotional responses following social exclusion, presents a valuable opportunity for studying reputation and reputation management in controlled environments. Participants in Cyberball can experience inclusion, exclusion, or over-inclusion from virtual peers, triggering social distress and activating related brain processes even when participants are aware of the programmed nature of the exclusion (Bolling et al., 2011; Crowley et al., 2009;

Williams, 2007). This ability to elicit genuine emotional responses in a controlled setting makes Cyberball a useful tool for studying reputation and reputation management.

In conjunction with Cyberball, economic games can offer a structured approach to studying social behaviour and incorporate real outcomes, such as monetary bonus payments, which reflect genuine preferences and decisions. Unlike the hypothetical scenarios or self-reported measures primarily used in this thesis, economic games ensure decisions have real consequences, providing insights into underlying motives and behaviours (Fehr & Schmidt, 1999; Gächter & Fehr, 1999).

Integrating Cyberball with variations of economic games could provide a comprehensive understanding of how reputation cues influence social behaviour. For instance, participants could play a Trust Game before or after Cyberball, where they experience inclusion or exclusion. Researchers could manipulate reputation cues throughout to explore questions such as whether participants exclude defectors in Cyberball, depending on whether they directly experienced the defection or heard about it through gossip. They could also investigate whether participants preferentially toss the ball to those they believe will play with them in future economic games and how they behave in Cyberball after accidentally defecting in economic games. Additionally, researchers could examine whether participants attempt to mitigate over-inclusion in Cyberball during subsequent economic games.

Furthermore, using Cyberball to manipulate fear of social exclusion (Williams et al., 2000) offers a tool to investigating cross-cultural differences in reputational concerns. Given the cross-cultural variations in socialisation and behaviours that influence one's reputation (Fu et al., 2016), future studies could undertake comparative analyses of how different cultures perceive and manage reputational concerns. Since reputational concern is often linked to fear of social exclusion (Cavazza et al., 2014; Pagliaro et al., 2016), I would expect to observe culturally-mediated differences in the degree of concern regarding judgments from strangers versus those within one's in-group (Uskul & Over, 2017).

### **7.5.3 Reputation Management and (Avoidance of) Embarrassment**

Future research could explore how emotions such as embarrassment or shame influence reputation management. One approach could involve using experience sampling methods to capture daily reflections on recent moments of embarrassment or attempts to avoid it. This

could provide real-world examples of reputation management, as participants report instances of embarrassment and their pre-emptive or subsequent behaviours.

Expressing negative emotions such as embarrassment following a transgression can aid in reputational recovery (Feinberg et al., 2012; Goffman, 1956). It signals prosociality and commitment to social relationships, with observers more willing to give resources and express a desire to affiliate with people expressing embarrassment (Feinberg et al., 2012).

People might feel embarrassment to (i) prevent making choices that could lead to devaluation, (ii) stop negative information from reaching others, and (iii) minimise the adverse effects of devaluation when it occurs (Sznycer et al., 2016). Essentially, embarrassment may serve to mitigate the likelihood or costs of reputation-damaging information spreading to others.

Therefore, behaviours that elicit more negative evaluations from others should elicit more embarrassment (Sznycer et al., 2016). Factors such as the intentionality of social norm violations can impact the level of embarrassment, with unintentional violations being less embarrassing than intentional ones (Bas-Hoogendam et al., 2018).

Future studies could examine the strategies employed to either restore or prevent a negative image. Researchers could also investigate situations in which individuals hide versus showcase feelings of embarrassment—likely showcasing when attempting to restore their image and hiding when trying to prevent a negative image. Lastly, comparisons could be made between reputation management strategies employed when feeling shame versus embarrassment. People feel shame when they believe a real flaw of the self has been revealed, and feel embarrassed when they believe others have reason to think a flaw has been revealed; embarrassment is tied to violations of conventions, while shame is tied to moral failings (Sabini et al., 2001).

#### **7.5.4 Reputation and Individual Differences**

Lastly, individual differences can influence people's concern about their reputation (Cavazza et al., 2015). Chapter 2, for instance, demonstrated that people who scored high on agreeableness, as well as on social interaction anxiety and social phobia scales, were more concerned about their reputation. Future studies could investigate whether various other individual differences, such as paranoia, affect how often individuals consider how others perceive them and which cues (e.g., status, groups) potentially mediate this effect. Mild paranoia is relatively common in the general population (e.g., Freeman et al., 2005) and includes features such as self-

consciousness, social fear, and discomfort with social interactions (Horton et al., 2014). Thus, a potential study might first determine where participants lie on the paranoia scale and then measure how often they experience reputational concerns. Further questions could assess the degree of reputational threat felt in each instance, as well as details of who was involved (including power structure, ingroup/outgroup perception, and whether it concerned an individual or group). As it has been suggested that more paranoid individuals cooperate less and punish more in dictator games (Raihani & Bell, 2018), one could also investigate instances in which these individuals cooperated or punished strategically to manage their reputation. By exploring various individual differences, researchers could gain a deeper understanding of the factors that shape reputation management strategies across different trait or cognitive ability profiles.

### **7.5.5 Research Linked to Open-Text Data**

Another fruitful avenue for further study could involve leveraging the richness of the open-text data collected in this thesis, particularly from Chapters 2 (reputation management events), 4 (cooperative and uncooperative behaviours), and 5 (third- and second-party punishing actions), using methods such as text mining or sentiment analysis.

For example, the free-text responses provided by participants in Chapter 2 could also be analysed to identify common methods of reputation management (e.g., lying, hiding behaviours, advertising actions, adjusting behaviours) and the domains in which reputation management occurred (e.g., prosociality, competence, attractiveness, norm compliance). Additionally, these entries could be coded based on whether participants were attempting to avoid negative reputations or gain positive ones, enabling a comparison of reputational concerns between these two motivations. Such an investigation may reveal that participants focus more on avoiding negative reputations than on gaining positive ones. This prediction is based on the difficulty with which good reputations are constructed and the ease with which they are lost (Cavazza et al., 2015; Nichols & Fournier, 1999; Yaniv & Kleinberger, 2000), the slow-to-change or "sticky" nature of reputations (e.g., Bromley, 2001; Fombrun, 1996), and the manner in which reputations track patterns of behaviour over time (Zinko et al., 2012).

## **7.6 Conclusion**

This thesis has advanced our understanding of reputation in naturalistic settings, offering insights that bridge the gap between controlled laboratory environments and the complexities of everyday social interactions. By employing a variety of methodological approaches, including daily diaries, longitudinal studies, and vignette experiments, this research has demonstrated that while fundamental reputation mechanisms observed in controlled settings often hold true in real-world contexts, there are nuances and context-dependent variations that must be considered.

The findings demonstrate that reputations are influenced by social relationships, situational contexts, and perceived motivations behind actions. My research particularly underscores the significance of motive attributions in shaping reputational outcomes. Whether in cooperative or punitive contexts, the perceived motives behind actions play a crucial role in how individuals are judged by others. This has practical implications for various stakeholders, including individuals, organisations, and policymakers, who can leverage these insights to foster more cooperative or productive environments.

In summary, this thesis contributes to a more nuanced understanding of how reputations are managed and perceived in everyday life. It calls for a broader, more context-aware approach to studying human behaviour, one that appreciates the intricate interplay of social factors influencing reputation. Future research should continue to explore these dynamics, further bridging the gap between controlled experimental findings and the rich tapestry of human social interactions in the real world.



# Appendices

# **Appendix A**

## **Supplementary Information for Chapter 2**

### **A1 Preregistration for “Reputation Management in Everyday Life”**

## Reputation Management in Everyday Life (#73338)

### Author(s)

This pre-registration is currently anonymous to enable blind peer-review.  
It has one author.

Pre-registered on: 08/25/2021 01:46 AM (PT)

### 1) Have any data been collected for this study already?

No, no data have been collected for this study yet.

### 2) What's the main question being asked or hypothesis being tested in this study?

1) Participants will report higher levels of reputational concern (i) when interacting with well-connected others, (ii) when interacting with in-group versus out-group members, (iii) when they value and (iv) feel close to the person(s), (v) when future interactions are likely, and (vi) when they are in public. 2) When engaging in reputation management, participants will be more concerned about trying to avoid reputational damage than they will be about trying to gain reputational benefits. 3) Participants will report higher reputational concern when they are managing their reputation compared to when they are engaged in other recent social interactions.

### 3) Describe the key dependent variable(s) specifying how they will be measured.

All three hypotheses will use 'level of reputational concern felt' as dependent variable. This will be measured on a slider scale from 0 – 1 ("Not at all concerned" – "Extremely Concerned") with 0.1 increments.

### 4) How many and which conditions will participants be assigned to?

All participants will first be asked whether they recently managed their reputation. If participants have, they answer questions regarding that branch, if participants have not, they instead answer similar questions about a recent social interaction. Participants receive this survey seven times (once daily for a week), and can follow either branch each time.

### 5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

- 1) Mixed effects model with level of reputational concern as outcome variable. The above described factors connectivity (continuous, scale from 0-1 with 0.1 increments), relationship type (categorical), relationship value (continuous, scale from 0-1), closeness (continuous, scale from 0-1), likelihood of future interaction (continuous, scale from 0-1), and audience (categorical) will be set as fixed effects. Participant ID will be added as random effect with random intercepts.
- 2) Mixed effects model with reputational concern as outcome variable, and reputational motivation type (categorical) as fixed effect. Participant ID will be added as random effect with random intercepts.
- 3) Mixed effects model with reputational concern as outcome variable and branch (categorical) as fixed effect. Participant ID will be added as random effect with random intercepts.

### 6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

Removals will be based on free text entries which will act as bot, language, attention, and comprehension checks. Unsuitable responses will be removed.

### 7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

Based on limited funds, we will recruit 240 participants (200 participants + 20% attrition).

### 8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?) We

will also administer an intake questionnaire (incl. demographics, personality, social anxiety, perspective taking questionnaire etc) with which we may want to look into secondary hypotheses, e.g.:

- Those with higher levels of social interaction anxiety/phobia (SIAS-6 & SPS-6) will experience more concern for reputation
- Political conservatives (SECS) may be more concerned with how they are viewed by family and close friends relative to liberals, while political liberals may be more concerned with how they appear to more distant connections than conservatives
- We may want to explore whether reputational concern varies with personality (Big Five Inventory) - We may also want to look into age/gender/ethnicity effects

## A2 Study Materials

### A2.1 Reputation Management in Everyday Life Qualtrics Intake Survey

---

#### Start of Block: Info and consent

##### Info

*This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A. The purpose of this study is to investigate people's reputation management in their everyday lives. **Please read the following information carefully in order to successfully complete the study.***

##### **Earn up to £6.25 !**

*The study involves an initial intake questionnaire, as well as the installation and use of an app (PIEL Survey) for a week. You will not be compensated for the initial questionnaire without taking part in the scheduled surveys as well (more details below). Please only take part if you generally have access to your smartphone for around 5min at some point between 7pm and 9pm.*

This study is being run by Nicole Engeler ([nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk)) under the supervision of Professor Nichola Raihani ([n.raihani@ucl.ac.uk](mailto:n.raihani@ucl.ac.uk)) in the Department of Experimental Psychology at UCL.

In this study, we will assess your everyday reputation management using the PIEL (Participation in Everyday Life) Survey app which will present you with **approx. 4-minute** long surveys **once a day for seven consecutive days**. The PIEL Survey app is free, does not contain any ads, and does not record any personal information about you. You will receive in-app notifications to complete the survey at **19:00 o'clock** (your local time) every day. The app will send a reminder 30min, 1h and 1h30min after the initial notification. Once you have opened a survey, all remaining reminders are deleted. **After 2h the survey will expire**, and you will no longer be able to complete that survey. If you miss a survey, do not worry, the schedule will continue, and you can answer the next survey. Just answer as many surveys as you can. You will receive **45p** for each completed survey. You will also receive a **bonus of £1** if you complete at least 80% of the scheduled surveys.

In the scheduled surveys we will first ask you **whether you have recently managed your reputation**, i.e. whether you have engaged in intentional actions to create or maintain a desired image of yourself in others – in other words, whether you have influenced the way others perceive you. We will ask you to report on a specific instance. Subsequently, we will ask questions about the situational information around that specific instance (e.g. how close you feel to the person(s) whose view of yourself you wanted to influence). Whenever you have not managed your reputation, we will ask you to answer similar questions about your most recent social interaction instead. In some instances, you will be asked to provide text responses about your experiences - please use initials or pseudonyms whenever you talk about others (or yourself) in these responses.

**Before the scheduled surveys can begin, you will first have to install the PIEL Survey app onto your smartphone, and subsequently import our survey file into the app.** We provide instructions to help with this process, however, please email us at [nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk) or message us on Prolific

if you are having troubles.

Prior to beginning the daily surveys, you will also be asked to fill out a battery of questionnaires about your demographic information (e.g. gender, age, education level), concern for reputation, personality, political ideology, and attitudes towards social interactions. This should take approx. **10 minutes** to complete. You will be paid **£2** for completing this initial questionnaire. However, you will not receive any compensation for completing this intake questionnaire unless you complete at least 3 scheduled surveys. The scheduled surveys **must be started within three** days of taking the intake questionnaire (which will immediately follow this introduction). You will receive an **additional bonus of 10p** if you install and activate our scheduled survey **within 24h** of completing the intake questionnaire (as proof, please send us a **code word** within those 24h). We will provide more detailed instructions on this following the initial questionnaire.

At the end of the study the app will prompt you to **email your data to us** (don't worry, it's very easy). We recommend that you do not delete your app or data until your submission has been approved in case there have been any problems with receiving your data, without which we cannot pay you.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous (your name will not appear in any publication related to this study and will not be shared with any other parties). Only researchers working with the Raihani Lab will have access to datafiles containing identifying information (Prolific ID etc). Completely de-identified datafiles (e.g. no Prolific ID) will be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. Before sending your data to us at the end of the study, you can delete your data at any time, either within the app or by deleting the app. Note that we unfortunately cannot compensate you for surveys you have completed if you delete your responses, as we have no way to assess how many surveys were completed.

Once de-identified data have been shared, it may be difficult or impossible to withdraw consent for your anonymous data to be used. **Please only take part in this study if you are happy for your anonymous data to be used in this way.**

#### **Reward Breakdown:**

- £2 for the Intake Survey (now)
- £3.15 for completing all 7 scheduled surveys (45p per survey)
- £1 for completing 80%+ of the scheduled surveys
- 10p for activating the scheduled surveys within 24h

---

Page Break

#### **notice Local Data Protection Privacy Notice**

##### **Notice:**

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

This 'local' privacy notice sets out the information that applies to this particular study. Further information on how UCL uses participant information can be found in our 'general' privacy notice:

For participants in health and care research studies, click [here](#)

The information that is required to be provided to participants under data protection legislation (GDPR and DPA 2018) is provided across both the 'local' and 'general' privacy notices.

The lawful basis that will be used to process your personal data are: 'Public task' for personal data and 'Research purposes' for special category data.

Your personal data will be processed so long as it is required for the research project. If we are able to anonymise or pseudonymise the personal data you provide we will undertake this, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return your submission and either delete the app, delete the data or simply don't send your collected responses.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)

Data Protection Officer: Alex Potts, [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)



Consent Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can email the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
- (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
- (4) I understand that my anonymized research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).

(5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning my submission.

- ☐ I **consent** to take part in this study. (1)
- ☐ I do **NOT consent** to take part in this study. (0)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent </strong>to take part in this study.*

End of Block: Info and consent

---

Start of Block: Demographics



PID Please enter your Prolific ID:

---

---

gender What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



age What is your age in years?

---



ethn What is your ethnic group?

- ☐ Asian (1)
  - ☐ Black (2)
  - ☐ Mixed (3)
  - ☐ Middle Eastern (4)
  - ☐ White (5)
  - ☐ Other (6) \_\_\_\_\_
  - ☐ Prefer not to say (7)
- 

nation What is your nationality?

\_\_\_\_\_

---

educ What is the highest level of education you have obtained? If you are foreign or your qualification is not listed, please tick the box that contains its nearest equivalent or select "other".

- ☐ GCSEs (1)
  - ☐ A-levels (2)
  - ☐ Apprenticeship (3)
  - ☐ Professional qualification (e.g. teaching, nursing) (4)
  - ☐ Bachelor Degree (5)
  - ☐ Higher Degree (e.g. MA, PhD, PGCE) (6)
  - ☐ Other (7) \_\_\_\_\_
-



income What is your total combined household income for the past 12 months, before taxes, from all sources, wages, public assistance/benefits, help from relatives, alimony, and so on? If you don't know your exact income, please estimate.

- ☐ Less than £9,999 (1)
- ☐ £10,000 - £19,999 (2)
- ☐ £20,000 - £49,999 (3)
- ☐ £50,000 - £99,999 (4)
- ☐ £100,000 - £149,999 (5)
- ☐ £150,000 - £174,999 (6)
- ☐ More than £175,000 (7)
- ☐ Prefer not to say (8)

#### End of Block: Demographics

#### Start of Block: Reputational Concern

rep Please indicate how characteristic the following statements are of you. Responses range from "Not at all characteristic of me" to "Extremely characteristic of me".

|                                                             | Not at all<br>characteristic<br>for me (1) | Slightly<br>characteristic<br>for me (2) | Moderately<br>characteristic<br>for me (3) | Very<br>characteristic<br>for me (4) | Extremely<br>characteristic<br>for me (5) |
|-------------------------------------------------------------|--------------------------------------------|------------------------------------------|--------------------------------------------|--------------------------------------|-------------------------------------------|
| I am rarely<br>concerned<br>about my<br>reputation. (1)     | <input type="radio"/>                      | <input type="radio"/>                    | <input type="radio"/>                      | <input type="radio"/>                | <input type="radio"/>                     |
| I do not<br>consider what<br>others say<br>about me. (2)    | <input type="radio"/>                      | <input type="radio"/>                    | <input type="radio"/>                      | <input type="radio"/>                | <input type="radio"/>                     |
| I wish to have a<br>good<br>reputation. (3)                 | <input type="radio"/>                      | <input type="radio"/>                    | <input type="radio"/>                      | <input type="radio"/>                | <input type="radio"/>                     |
| If my<br>reputation is<br>not good, I feel<br>very bad. (4) | <input type="radio"/>                      | <input type="radio"/>                    | <input type="radio"/>                      | <input type="radio"/>                | <input type="radio"/>                     |
| I find it<br>important that<br>others consider              | <input type="radio"/>                      | <input type="radio"/>                    | <input type="radio"/>                      | <input type="radio"/>                | <input type="radio"/>                     |

my reputation  
as a serious  
matter. (5)

I try hard to  
work on my  
reputation (in  
my  
relationships  
with others).  
(6)

I find it difficult  
if others paint  
an incorrect  
image of me.  
(7)

☐ ☐ ☐ ☐ ☐
☐ ☐ ☐ ☐ ☐

#### End of Block: Reputational Concern

#### Start of Block: Big 5

big5 Here are a number of characteristics that may or may not apply to you. Please indicate the strength of your agreement with each statement by ticking the appropriate box. Responses range from "disagree strongly" to "agree strongly".

I see myself as someone who...

|                                                | Disagree<br>strongly (1) | Disagree a little<br>(2) | Neither agree<br>nor disagree<br>(3) | Agree a little<br>(4) | Agree strongly<br>(5) |
|------------------------------------------------|--------------------------|--------------------------|--------------------------------------|-----------------------|-----------------------|
| Is talkative (1)                               | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| Tends to find<br>fault with<br>others (2)      | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| Does a<br>thorough job (3)                     | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| Is depressed,<br>blue (4)                      | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| Is original,<br>comes up with<br>new ideas (5) | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |

|                                                   |                       |                       |                       |                       |                       |
|---------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Is reserved (6)                                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is helpful and<br>unselfish with<br>others (7)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Can be<br>somewhat<br>careless (8)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is relaxed,<br>handles stress<br>well (9)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is curious about<br>many different<br>things (10) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is full of energy<br>(11)                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Starts quarrels<br>with others (12)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is a reliable<br>worker (13)                      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Can be tense<br>(14)                              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is ingenious, a<br>deep thinker<br>(15)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Generates a lot<br>of enthusiasm<br>(16)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Has a forgiving<br>nature (17)                    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Tends to be<br>disorganized<br>(18)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Worries a lot<br>(19)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                                                 |                       |                       |                       |                       |                       |
|-------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Has an active imagination (20)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Tends to be quiet (21)                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is generally trusting (22)                      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Tends to be lazy (23)                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is emotionally stable, not easily upset (24)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is inventive (25)                               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Has an assertive personality (26)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Can be cold and aloof (27)                      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Perseveres until the task is finished (28)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Can be moody (29)                               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Values artistic, aesthetic experiences (30)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is sometimes shy, inhibited (31)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is considerate and kind to almost everyone (32) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Does things efficiently (33)                    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                                                    |                       |                       |                       |                       |                       |
|----------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Remains calm in tense situations (34)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Prefers work that is routine (35)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is outgoing, sociable (36)                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is sometimes rude to others (37)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Makes plans and follows through with them (38)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Gets nervous easily (39)                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Likes to reflect, play with ideas (40)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Has few artistic interests (41)                    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Likes to cooperate with others (42)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is easily distracted (43)                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Is sophisticated in art, music, or literature (44) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: Big 5

Start of Block: Social and Economic Conservatism Scale (SECS)



SECS Please indicate the extent to which you feel positive or negative towards each issue. Scores of 0 indicate greater negativity, and scores of 100 indicate greater positivity. Scores of 50 indicate that you feel neutral about the issue.



End of Block: Social and Economic Conservatism Scale (SECS)

Start of Block: SIAS-6 and SPS-6



soc anx For each question, please indicate the degree to which you feel the statement is characteristic or true of you. Responses range from "Not at all characteristic or true of me" to "Extremely characteristic or true of me".

|  |                              |                            |                              |                        |                             |
|--|------------------------------|----------------------------|------------------------------|------------------------|-----------------------------|
|  | Not at all<br>characteristic | Slightly<br>characteristic | Moderately<br>characteristic | Very<br>characteristic | Extremely<br>characteristic |
|--|------------------------------|----------------------------|------------------------------|------------------------|-----------------------------|

|                                                                            | or true of me<br>(1)  | or true of me<br>(2)  | or true of me<br>(3)  | or true of me<br>(4)  | or true of me<br>(5)  |
|----------------------------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| I have difficulty making eye contact with others. (1)                      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I find it difficult mixing comfortably with the people I work with. (2)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I tense up if I meet an acquaintance on the street. (3)                    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I feel tense if I am alone with just one person. (4)                       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I have difficulty talking with other people. (5)                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I find it difficult to disagree with another's point of view. (6)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I get nervous that people are staring at me as I walk down the street. (7) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I worry about shaking or trembling when I'm watched by other people. (8)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

I would get tense if I had to sit facing other people on a bus or train. (9)

|                       |                       |                       |                       |                       |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|

I worry I might do something to attract the attention of other people. (10)

|                       |                       |                       |                       |                       |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|

When in an elevator, I am tense if people look at me. (11)

|                       |                       |                       |                       |                       |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|

I can feel conspicuous standing in a line. (12)

|                       |                       |                       |                       |                       |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|

End of Block: SIAS-6 and SPS-6

Start of Block: SoPerTa



socPerTaking

Please indicate how often you engage in the following behaviours. Responses range from "Almost never" to "Almost all the time".

|                                                                                                               | Almost never<br>(1)   | Rarely (2)            | Sometimes (3)         | Often (4)             | Almost all the<br>time (5) |
|---------------------------------------------------------------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------------------|
| How often do you try to figure out how the people around you view different situations? (1)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>      |
| If you are having a disagreement with your friends, how often do you try to imagine how they are feeling? (4) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>      |
| How often do you try to look at everybody's side of a disagreement                                            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>      |



|                                                                                                                                     |                       |                       |                       |                       |                       |
|-------------------------------------------------------------------------------------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| before you make a decision? (5)                                                                                                     |                       |                       |                       |                       |                       |
| When you are upset at someone, how often do you try to “put yourself in his or her shoes”? (6)                                      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| How often do you try to understand your friends/co-workers/acquaintances better by trying to figure out what they are thinking? (7) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Before criticizing others, how often do you imagine how you would feel if you were in their place? (8)                              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| To understand your friends better, how often do you imagine how things look from their perspective? (9)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: SoPerTa

Start of Block: detail

detail

### IMPORTANT

*It is vital that you carefully read the following so that you are able to successfully complete the scheduled surveys.*

In the scheduled PIEL surveys we will ask you whether you have recently managed your reputation, i.e. whether you have engaged in intentional actions to create or maintain a desired image of yourself in others. We will ask you to report on a specific instance. As it is not always easy to remember or notice when you manage your reputation, a tip is to first think about your recent social interactions (or future interactions that you are anticipating) and to then reflect whether you engaged in reputation management at any point.

As we manage our reputations in many diverse ways, the surveys aim to gain an understanding about what you (don't) do *to influence someone's perception about you*. Therefore, please help us by writing sentences that are comprehensible to others who were not in your situation when you are provided with free text entry boxes (e.g. by providing enough context and fully formulating your thoughts). Here are some examples:

- 1) "I opened the blinds really early so that the neighbour didn't think I am lazy and sleep in late."
- 2) "I had my zoom camera off so that nobody would see I'm texting during the presentation. I didn't want people to know I wasn't paying attention."
- 3) "I spent a fortune at the supermarket and cooked a big meal for my friends and insisted they don't pay me anything back. I want them to think of me as a great hostess."
- 4) "I was out of breath when I got home and held my breath when I saw my roommate. I didn't want her to know I'm in such bad shape."
- 5) "I hid the fact that I was playing Pokemon go when walking around so people didn't think I was a loser."

Thank you very much for agreeing to take part in this study! If you have any questions, please do not hesitate to contact us on Prolific or at [nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk)

End of Block: detail

---

Start of Block: PIEL Instruction Link

PIEL

Thank you for completing the intake questionnaire!

We will now link you to the '[PIEL Survey App Instructions](#)' where we will provide you with information on how to install and use the app, including how to load our survey file, how to send us your data at the end, and where to find the code word. Please read those instructions carefully so that you do not miss anything important.

Please make sure to open the [PIEL Survey App Instructions](#) before moving on, as you will be redirected back to Prolific.

In case you wish to manually copy the PIEL Survey App Instructions link, please copy the following:  
**<https://osf.io/7w5mn/>**

End of Block: PIEL Instruction Link

---

## A2.2 Scheduled Daily Surveys as Loaded into the PIEL App

```
# 1x for 7 days (7pm); reminders after 30min, 1h, 1h30min; 2h response window; scheduled; cannot test; shows % of past surveys completed; whether managed own reputation (otherwise recent social interactions)
```

```
name|'Reputation management in everyday life'
```

```
author-email|nicole.engeler.19@ucl.ac.uk
```

```
author|Nicole Engeler
```

```
max-delay|7200
```

```
can-run-once|0
```

```
run-for|7
```

```
run|*|1900
```

```
reminders|1800,3600,5400
```

```
can-test|0
```

```
exit-message|4|Thank you for taking part!
```

```
placeholder-message|No questions for now, thank you for checking! The code word is: Hufflepuff
```

```
compliance-display|2
```

```
1|'Reputation management' is when we engage in actions that will create or maintain a desired image of ourselves in others. Please try to think of a recent instance in which you have managed your reputation. %HINT It helps to think about recent social interactions! %TYPE list |OK
```

```
2|Have you tried to make others perceive you in a certain way since the last survey? %HINT %TYPE list|Yes|No %NEXT 13
```

```
3|How did you manage your reputation (i.e. what did you do, what was the situation)? %HINT Please write a few sentences. %TYPE text
```

```
4|Why did you manage your reputation (i.e. what were your motivations and feelings)? %HINT In what way did you want to be perceived? %TYPE text
```

```
5|How concerned did you feel for how you might appear to the other person(s)? %HINT %TYPE slider|Not at all concerned|Extremely concerned
```

```
6|How much do you value the relationship with the person(s) whose perception you wanted to manage? %HINT %TYPE slider|Not at all|Very much
```

```
7|At this moment, how close do you feel to the person(s) whose view of yourself you wanted to influence? %HINT %TYPE slider|Not at all close|Very close
```

8|What relationship do you have with the person(s)? %HINT %TYPE  
list|Stranger %NEXT 10|Acquaintance %NEXT 10|Colleague/ Classmate %NEXT  
10|Instructor/ Superior %NEXT 10|Friend %NEXT 10|Romantic Partner %NEXT  
10|Family %NEXT 10|Other

9|Other: %HINT %TYPE text

10|How well-connected is the person(s) to others who know you? %HINT Do  
you know many of the same people? %TYPE slider|Not at all  
connected|Very connected

11|How likely is it that you will interact with the person(s) in  
future? %HINT %TYPE slider |Not at all likely|Very likely

12|Did you manage your reputation while you were alone and unobserved,  
or in front of others (this includes the person(s) whose perception you  
were managing)? %HINT %TYPE list |Alone %NEXT END|In front of others  
%NEXT END

13|What did your most recent social interaction look like (i.e. what  
did you do, what was the situation)? %HINT %TYPE text

14|Why did you act the way you did during the social interaction (i.e.  
what were your motivations and feelings)? %HINT %TYPE text

15|During this interaction, how concerned did you feel for how you  
might appear to the other person(s)? %HINT %TYPE slider|Not at all  
concerned|Extremely concerned

16|How much do you value the relationship with the person(s) you  
interacted with? %HINT %TYPE slider|Not at all|Very much

17|At this moment, how close do you feel to the person(s) you  
interacted with? %HINT %TYPE slider|Not at all close|Very close

18|What relationship do you have with the person(s)? %HINT %TYPE  
list|Stranger %NEXT 20|Acquaintance %NEXT 20|Colleague/ Classmate %NEXT  
20|Instructor/ Superior %NEXT 20|Friend %NEXT 20|Romantic Partner %NEXT  
20|Family %NEXT 20|Other

19|Other: %HINT %TYPE text

20|How well-connected is the person(s) to others who know you? %HINT Do  
you know many of the same people? %TYPE slider|Not at all  
connected|Very connected

21|How likely is it that you will interact with the person(s) in  
future? %HINT %TYPE slider |Not at all likely |Very likely

## A2.3 PIEL Survey App Instructions

Thank you for taking part in our study about reputation management in everyday life. The survey is to be conducted using the PIEL Survey app. You should find it intuitive to use but some important information is provided here to make sure that you do not have any problems.

It is important that you do not stop the survey or delete the PIEL Survey app unless we confirm that you can do so at the end of the study, or if you wish to terminate the study (if you do so, all of your collected data will be deleted).

**When you are asked for your Participant ID, please input your Prolific ID.**

If you have questions or problems, please message us on Prolific or email us at [nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk)

### SURVEY SETUP

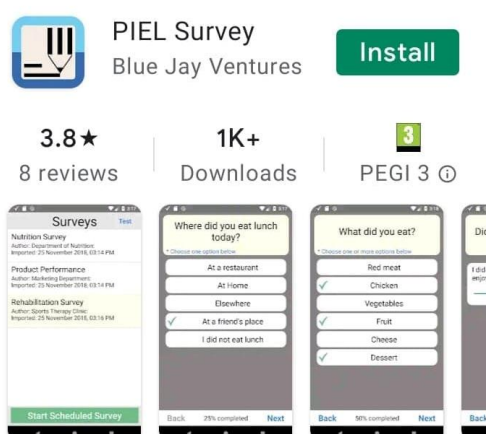
***You must install the PIEL Survey app on your smartphone before you are able to import the survey into the app. Everything must be done on your smartphone.***

***At the end of the week, it is up to you to send your data to us – please read these instructions carefully so that you do not miss anything important.***

### Download the app

The app “PIEL Survey” is available on the App store/ Google Play, depending on your phone type. Please use the relevant store to search for the app on your mobile device. Then install the PIEL Survey app. It is free.

Picture of the PIEL Survey app in Google Play



The PIEL Survey gathers research survey data from people in their daily lives.

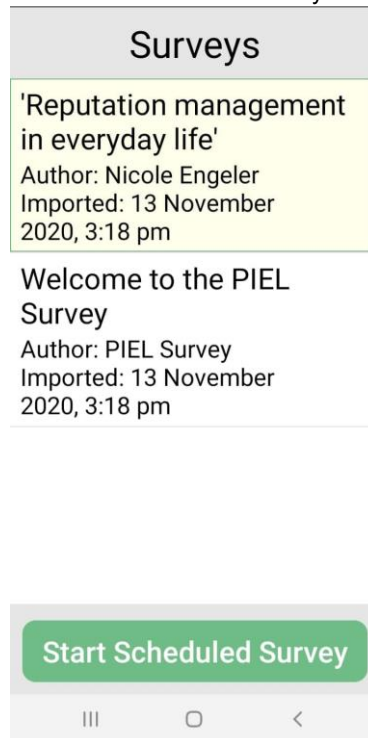
Once installed, you will notice that a "Welcome" survey will appear in the app. This survey is not related to our research survey but is a sample from the app developer. You can tap on it and take their survey as a way of exploring the app or delete it if you wish.

## Import Survey

Please download the file “ReputationManagementInEverydayLife.survey”, which you can access [here](#). You can open the survey in the app by tapping on the downloaded file and selecting the option to open it in the “PIEL Survey”.

Picture of the imported and selected survey

\*Note: some details such as date and time may differ on these images



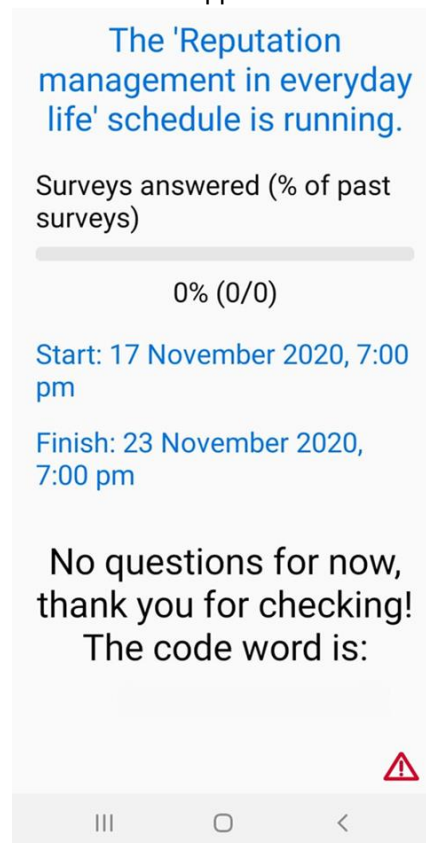
## Activate and Run Survey

Simply tap on the imported survey “Reputation Management in Everyday Life” and you will see that it is selected. Please tap the green button at the bottom of the screen (“Start Scheduled Survey”) to start the survey.

A placeholder screen will appear, with the name of the survey and the completion date. *This is also where you will see the **code word** with which you can prove that you have activated the scheduled surveys. Please either email this code word to [nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk) with your ID or message us on Prolific. You will receive a **10p bonus** if you send us the code word within 24h of taking the intake questionnaire.*

Picture of the placeholder screen

\*Note that the code word will appear at the bottom of the screen!



### Wait for survey

There is nothing to do for the moment, the app will notify you when a survey is due.

We ask that you do a few things:

- please do not disable notifications for the PIEL Survey app while a survey is running
- if you enable "Do not disturb", please exclude the PIEL Survey app if possible or remember to turn off "Do not disturb" when no longer needed
- please keep your device charged and nearby with the volume/ vibration at a level that you can easily hear

### Note: Android phones

Owners of some Android phones may require further steps to allow notifications to be received. If you are not receiving notifications, please see some examples on the [support website](#).

## COMPLETING A SURVEY

### Notification

The survey will run once a day for 7 days. You will receive in-app notifications to complete the survey at 19:00 o'clock (your local time) every day.

When you receive a notification, you can tap on the notification to start the survey or simply open the app.

If for any reason you cannot complete the survey immediately, the app will send a reminder 30min, 1h and 1h30min after the initial notification. After 2h the survey will expire, and you will no longer be able to complete that survey. Once you open a survey the remaining reminders will be deleted. If you miss a survey, don't worry, the schedule will continue, and you can answer the next survey. Just answer as many surveys as you can.

### Answering questions

Answering questions is easy. You must answer a question to be able to move forward, either by swiping or by tapping "Next". *If you close the app, you will terminate the current survey.*

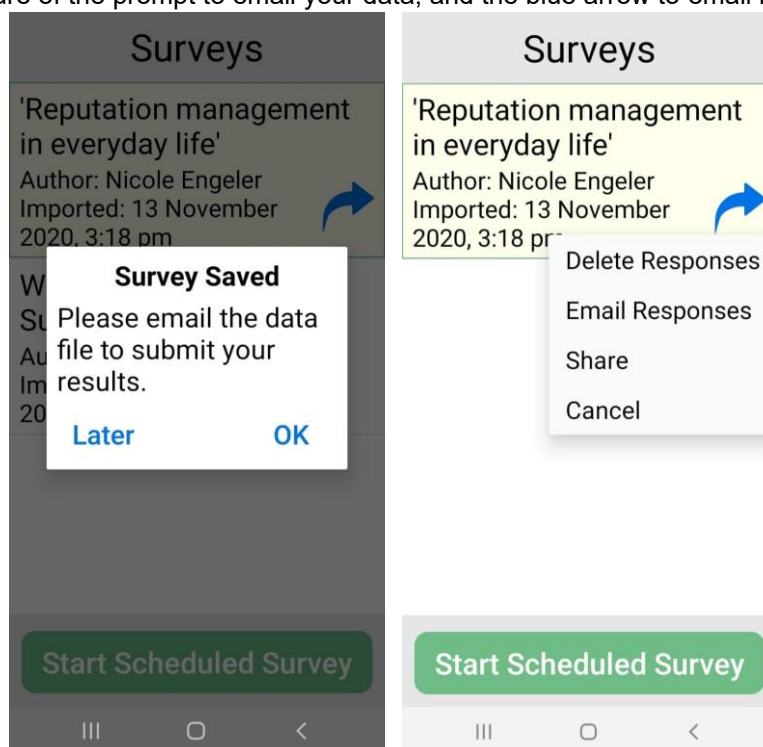
### Sending Data

The PIEL Survey app does not ever send out your data during a project. You are in complete control of your data.

At the end of all the scheduled surveys, you will receive an alert asking if you want to send your data to us. Please choose to email the data immediately by selecting "OK". It will be sent using your own email account directly to us. If you do not immediately send the email it will not delete the data and you can resend it later by pressing the blue arrow.

While this approach provides a high level of protection of your data, it is important that you DO NOT delete the data or the app until we have confirmed that we have received your data, so that we can ensure you receive appropriate payment before you delete your copy. If you are having issues, DO NOT restart the scheduled survey, but contact us.

Picture of the prompt to email your data, and the blue arrow to email it later.



**Thank you!**



### A3 Free Text Responses

Within this section, a random assortment of participant responses is showcased, to provide a brief overview of the diverse range of behaviours and desired reputations reported by participants. Alongside a spectrum of reputation management tactics, including engaging in specific actions (*"I cleaned my girlfriend's car for the first time to appear more helpful (...) I want to be perceived as a better person."*) or refraining from certain behaviours (*"I refused to take painkillers for a hurting knee to seem tougher to my fiancée."*), participants aimed to uphold and cultivate favourable reputations, or avoid unfavourable ones (e.g. *"I am a smoker and while being a guest at friends house I refrained from smoking to not be perceived by their parents as someone who is a bad influence"*). Reputation management endeavours extended across various domains, encompassing intelligence (*"I finished a difficult crossword puzzle on a plane really quickly so that that the people around me would think I was smarter than average."*), physical prowess (*"I got tired quickly at the gym, so I started limping so that people think I am injured lol. I was embarrassed that I got tired so quickly on the treadmill."*), and generosity (*"I am not in the greatest financial situation but as my friend visited me today I still insisted on paying for everything that day to be perceived as a good host. I wanted to be perceived as a considerate and welcoming person and not tight-fisted"*). Some efforts at reputation management displayed strategic foresight, as seen in the response provided by a PhD student: *"I assisted a partner at the lab even though I was a bit tired. She is new at the lab and I didn't want her to think I can't do what she was asking. Also, she is closer to the professor than me so it would be useful to be on her good side"*. Overall, these diverse responses underscore the complexity of reputation management in everyday life.

# Appendix B

## Supplementary Information for Chapter 3

### B1 Study Materials (Qualtrics Survey)

---

#### Start of Block: Info and consent

info *This study has been approved by the UCL Ethics Board. Project ID: 3720-002.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, we will record your evaluations of other people's punishing behaviours.**

This is a longitudinal study, meaning that you will be sent the same survey every Wednesday for six weeks. Please try to complete all six surveys.

You will receive **£0.70 for each completed survey**. You can earn an additional **bonus payment of £0.50** for completing 5 surveys, or a **bonus payment of £0.60** for completing all 6 surveys.

In the surveys we will ask you **whether you have recently witnessed or learned about a punishment act**. Subsequently, we will ask questions about the context (e.g., the severity of the violation being punished, the social relationships between those involved), and your evaluations of the situation and the punisher (e.g., whether the punishment was justified, what you think of the punisher).

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. Only researchers working with the Raihani Lab will have access to datafiles containing identifying information (e.g. Prolific ID). Completely de-identified datafiles (e.g. no Prolific ID) will be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return your submission.

---

Page Break

## notice **Local Data Protection Privacy Notice**

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: 'Public task'

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return your submission.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)



consent Please only consent to taking part in this study if you agree with ALL of the following statements:

(1) I have read the information above and understand I can email the researchers with any questions. (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing. (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications. (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared). (5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning my submission.

- ☐ I **consent** to take part in this study. (1)
- ☐ I do **NOT consent** to take part in this study. (0)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do **NOT consent** to take part in this study.*



PID Please enter your Prolific ID

---

End of Block: Info and consent

---

Start of Block: Demographics

gender What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



age What is your age in years?

---



nation What is your nationality?

---



ethn What is your ethnic group?

- ☐ Asian (1)
  - ☐ Black (2)
  - ☐ Mixed (3)
  - ☐ Middle Eastern (4)
  - ☐ White (5)
  - ☐ Other (6) \_\_\_\_\_
  - ☐ Prefer not to say (7)
- 

educ

What is the highest level of education you have obtained? If your qualification is not listed, please tick the box that contains its nearest equivalent or select "other".

- ☐ GCSEs (1)
- ☐ A-levels (2)
- ☐ Apprenticeship (3)
- ☐ Professional qualification (e.g. teaching, nursing) (4)
- ☐ Bachelor Degree (5)
- ☐ Higher Degree (e.g. MA, PhD, PGCE) (6)
- ☐ Other (7) \_\_\_\_\_

End of Block: Demographics

---

Start of Block: task info

info1 In our daily lives we punish people who violate social norms and expectations to varying degrees. Violations can be serious (e.g. stealing) or mundane (e.g. making an inconsiderate comment), and can be committed by someone familiar or a stranger.

---

Page Break

---

info2 Punishment is a behaviour that inflicts a cost on (or removes benefits from) a person in *response* to a specific behaviour by the person.

---

Page Break

---

info3 Punishment can take many forms, such as directly confronting a violator (either physically or verbally), sharing negative information about the violator to others (i.e., gossiping), demanding monetary compensation (i.e., material punishment), discipline (e.g., grounding a child), excluding the violator from social interactions or avoiding social contact with them.

---

Page Break

---

info4 The person who engaged in a wrong behaviour is the '*violator*'.

The *punisher* is the one who punishes the violator for their wrong behaviour.

---

Page Break

---

info5 Therefore, the violator *CANNOT* be the punisher, as the punisher punishes the violator in response to the violator's wrong behaviour.

---

Page Break

---

info6 The person who was harmed or wronged (*by the violator*) is the '*victim*'.

Therefore, the violator *CANNOT* be the victim.

But the punisher *can* be the victim (e.g., when the violator wrongs the punisher and the punisher punishes that violation - then the punisher is also the victim).

---

Page Break

info7 The victim could be you yourself, the punisher, another person, or even the general community such as when a park is being littered.

Note: even if the punishment harms the violator, the violator is NOT the victim – **the one who was affected by the violator's wrong behaviour is the victim.**

---

Page Break

info8 As we want to know **how you evaluate other people's punishing behaviours** in everyday life, **please do NOT report your OWN punishing behaviour.**

You can be the violator, the victim, or uninvolved. You cannot be the punisher.

---

Page Break

kp I understand that...

- ☐ Punishment occurs in response to a wrong behaviour (1)
- ☐ Punishment inflicts a cost on (or removes benefits from) the violator (2)
- ☐ The punisher is NOT the violator (3)
- ☐ The punisher *can* be the victim (4)
- ☐ The violator *cannot* be the victim (5)
- ☐ I should NOT report my own punishing behaviour (6)

Display This Question:

If I understand that... [q://QID54/SelectedChoicesCount Is Less Than 6](#)

repeat ***That was not quite right. All the previous statements were correct. Please read the following information again:***

In our daily lives we punish people who violate social norms and expectations to varying degrees. Violations can be serious (e.g. stealing) or mundane (e.g. making an inconsiderate comment), and can be committed by someone familiar or a stranger.

In this study, we consider punishment as a behaviour that inflicts a cost on (or removes benefits from) a person in response to a specific behaviour by that person.

Punishment can take many forms, such as directly confronting a violator (either physically or verbally), sharing negative information about the violator to others (i.e., gossiping), demanding monetary compensation (i.e., material punishment), discipline (e.g., grounding a child), excluding the violator from social interactions or avoiding social contact with them.

Within this survey we will refer to the person who engaged in the wrong behaviour as the 'violator'. We will refer to the person who was harmed or wronged as the 'victim'.

The victim could be you yourself, the punisher, another person, or even the general community such as when a park is being littered.

As this study aims to investigate your evaluations of someone else's punishing behaviour, please do NOT report your own punishing behaviour.

check1 Reminder: punishment is something that happens *in response* to a **wrong behaviour**.

check2

**Reminder**

The person who engaged in a wrong behaviour is the *violator*.



The punisher is the person who punishes the violator for engaging in a wrong behaviour.  
The victim is the person who was affected by the violator's wrong behaviour.

---

Page Break

example

**Example**

In a scenario where "my friends child swore at her, so she punished him by cancelling pizza night":

The violation is swearing, the punishment is cancelling pizza night (e.g., "withdrawal of benefits" and/or "discipline").

The **violation** is the **child**.

The **punisher** is the **parent**.

The **victim** is the **parent**.

Therefore, when asked who the victim is, you'd select "the punisher".

The **relationship** between the **punisher (parent)** and the **violation (child)** is "**family**".

For the **relationship** between **punisher** and **victim** you'd select "**the punisher was the victim**" as in this case the parent was both the punisher AND the victim.

---

Page Break

recent

In a moment, we will ask you whether you have witnessed or learned about a punishment act since the last survey.

This includes situations in which you yourself were punished, you watched someone else (e.g., a friend or a stranger) being punished in person or online, or you learned about a punishment act (e.g., because someone told or texted you about it, or you overheard gossip).

Please **focus on people** rather than country-level punishment.

Please take a moment to try to recall a recent punishment act. If multiple instances come to mind, **please choose only ONE punishment act to report on throughout this survey.**

*Remember: your OWN punishing behaviour does NOT count.*



event **Have you witnessed or learned about a punishment act since the last survey?** (Please stick to only ONE punishing instance throughout.)

Reminder: your OWN punishing behaviour does NOT count.

☐ Yes (1)

☐ No (0)

---

reminder

**The below is a reminder of what you have already read, in case you need it:**

In our daily lives we punish people who violate social norms and expectations to varying degrees. Violations can be serious (e.g. stealing) or mundane (e.g. making an inconsiderate comment), and can be committed by someone familiar or a stranger.

In this study, we consider **punishment as a behaviour that inflicts a cost on (or removes benefits from) a person in response to a specific behaviour by that person.**

Punishment can take many forms, such as directly confronting a violator (either physically or verbally), sharing negative information about the violator to others (i.e., gossiping), demanding monetary compensation (i.e., material punishment), discipline (e.g., grounding a child), excluding the violator from social interactions or avoiding social contact with them.

Within this survey we will refer to the person who engaged in the wrong behaviour as the 'violation'. We will refer to the person who was harmed or wronged by the violator as the 'victim'. The punisher is the one who punishes the violator for their wrong behaviour.

The victim could be you yourself, the punisher, another person, or even the general community such as when a park is being littered.

The punisher is not the violator. The violator is not the victim. The punisher can be the victim.

You can be the violator, the victim, or a bystander. You cannot be the punisher.

Witnessing or learning about punishment acts include situations in which you yourself were punished, you watched someone else (e.g., a friend or a stranger) being punished in person or online,

or you learned about a punishment act (e.g., because someone told or texted you about it, or you overheard gossip). Please focus on people rather than institution- or country-level punishment.

End of Block: task info

---

Start of Block: Punisher Evaluation

description Please briefly describe the *punishing act*.

**What action(s) did the PUNISHER take in response to the wrong behaviour?** Although it can be useful to mention what violation was punished (i.e., what the other person did to be punished), we are interested in the *punishing behaviour (i.e., how the violator was punished)*.

---

---

---

Page Break



form What form did the punishment take?

*If you select "Other", please expand.*

- ☐ Material (1)
- ☐ Avoidance (2)
- ☐ Exclusion (3)
- ☐ Withdrawal of benefits (4)
- ☐ Gossip (5)
- ☐ Verbal Confrontation (6)
- ☐ Physical Confrontation (7)
- ☐ Discipline (8)
- ☐ Other (9) \_\_\_\_\_

---

Page Break

severe How severe was the violation that was punished?

Not at all severe

Very severe



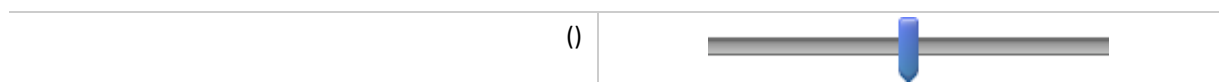
---

Page Break

just How justified was the punishment?

Not at all justified

Completely justified



Page Break

---



intent Do you believe the violation was accidental or intentional?

☐ Accidental (0)

☐ Intentional (1)

---

Page Break

---



violator Who was being punished (i.e., who was the violator)?

☐ Me (1)

☐ Someone else (0)

---

Page Break

---

victim **Who was the victim of the violation (i.e., who was affected by the violators wrong behaviour)?** Select all that apply.

☐ Me (1)

☐ The Punisher (2)

☐ Someone Else (3)

☐ The general community / environment (4)



violation relationship **What relationship do the *punisher* and the *violation* have to each other?**

The violator is the person(s) who engaged in the wrong behaviour that was punished.

- ☐ Strangers (1)
  - ☐ Acquaintances (2)
  - ☐ Colleagues/ Classmates (3)
  - ☐ Competitors (4)
  - ☐ Friends (5)
  - ☐ Romantic Partners (6)
  - ☐ Family (7)
  - ☐ Other (8) \_\_\_\_\_
- 



violation relationship **What relationship do the *punisher* and the *victim* have to each other?**

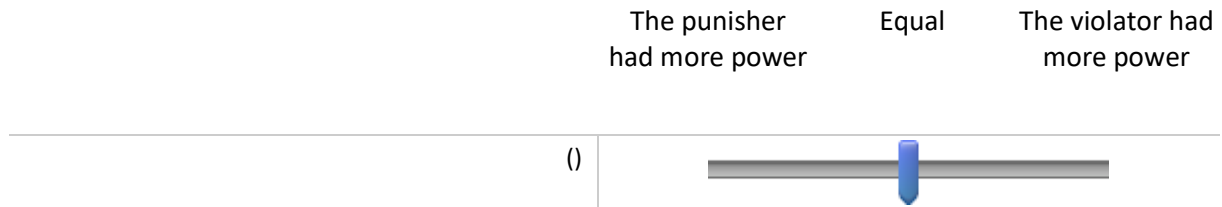
The victim is the person(s) affected by the violator's wrong behaviour.

**Note:** If you previously indicated that the punisher is the victim you should select the first option here.

- ☐ The punisher was the victim (0)
- ☐ Strangers (1)
- ☐ Acquaintances (2)
- ☐ Colleagues/ Classmates (3)
- ☐ Competitors (4)
- ☐ Friends (5)
- ☐ Romantic Partners (6)
- ☐ Family (7)
- ☐ Other (8) \_\_\_\_\_

Page Break

power Was there any power asymmetry between the punisher and the violator? This includes higher social status, dominance, prestige etc.



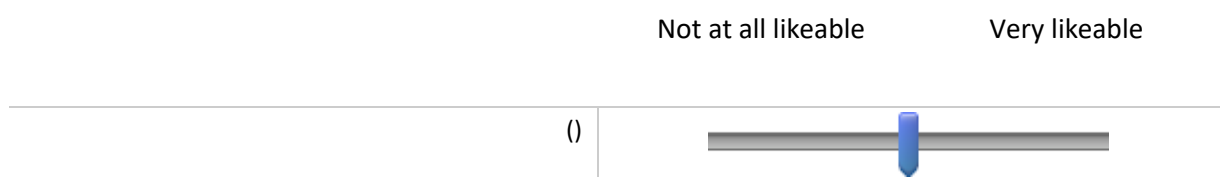
Page Break

motive Do you think that the punishers motivation for punishing was more altruistic or more self-serving?

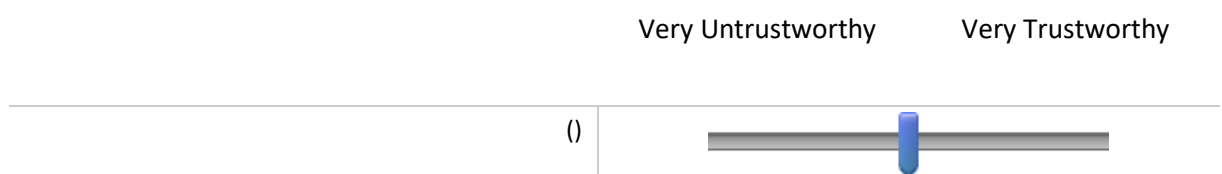


Page Break

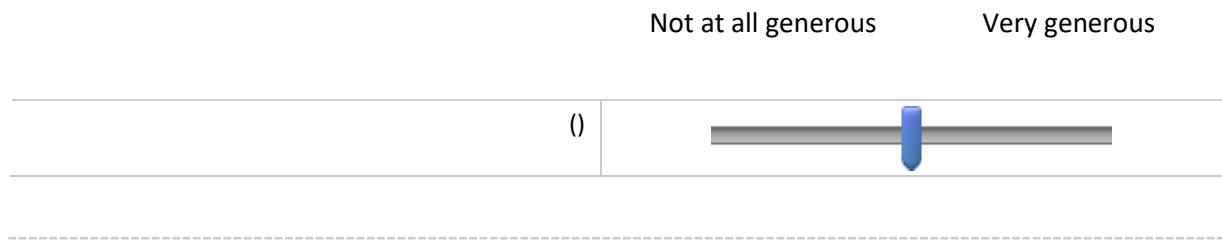
like Do you think the punisher is...



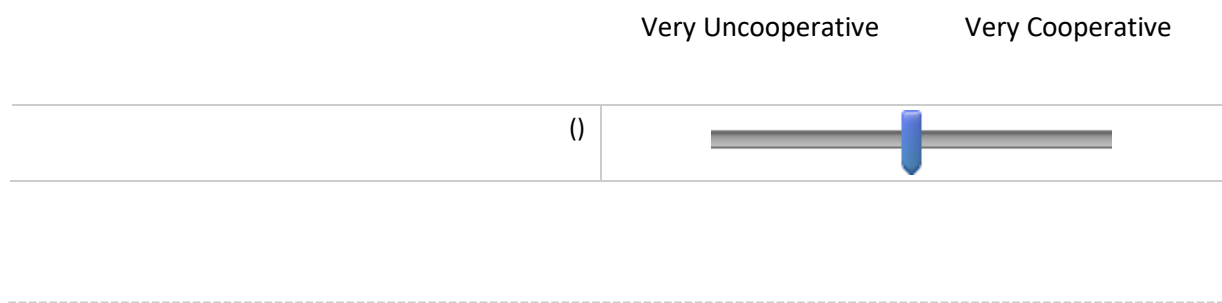
trust Do you think the punisher is...



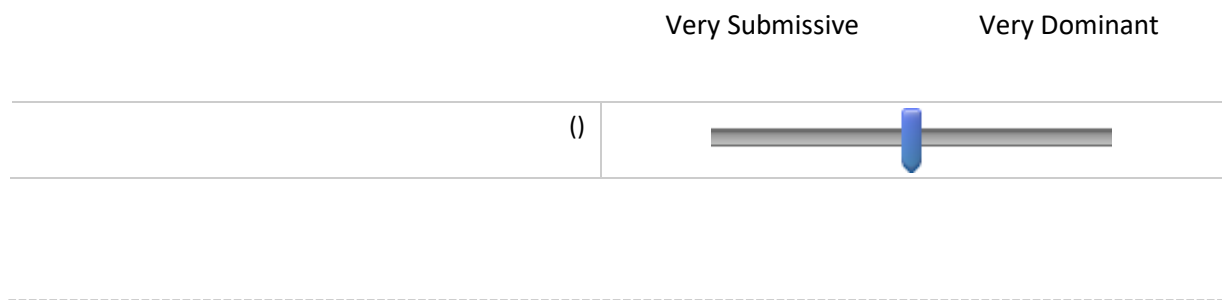
generous Do you think the punisher is...



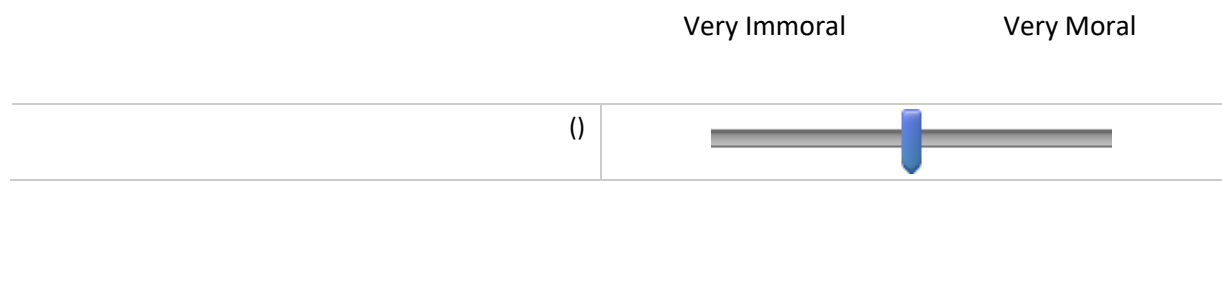
coop Do you think the punisher is...



subDom Do you think the punisher is...

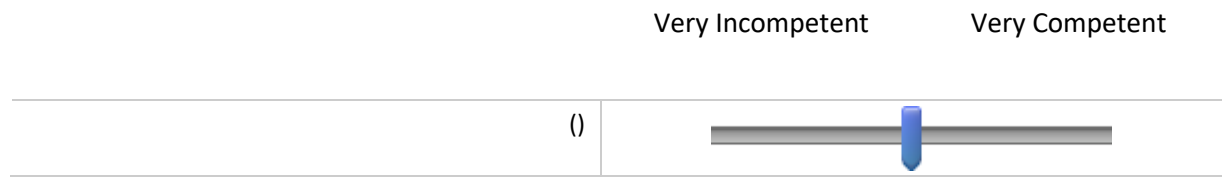


moral Do you think the punisher is...

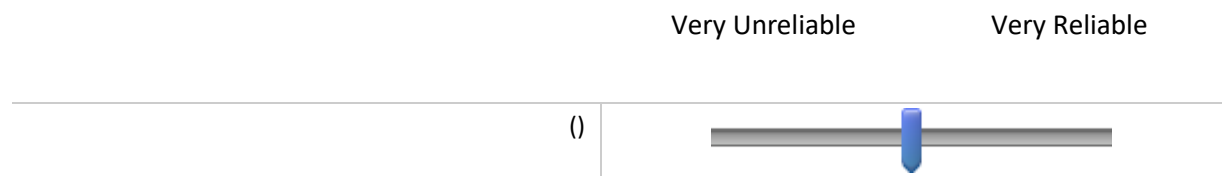




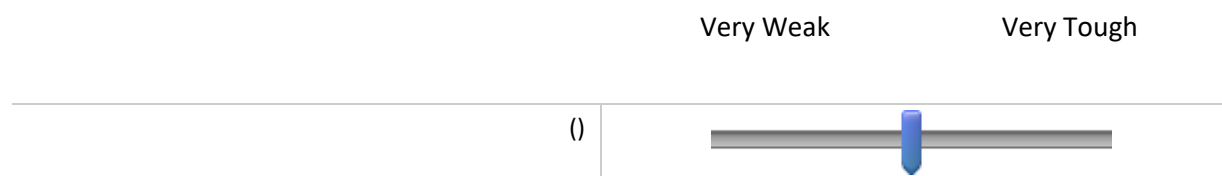
competent Do you think the punisher is...



reliable Do you think the punisher is...

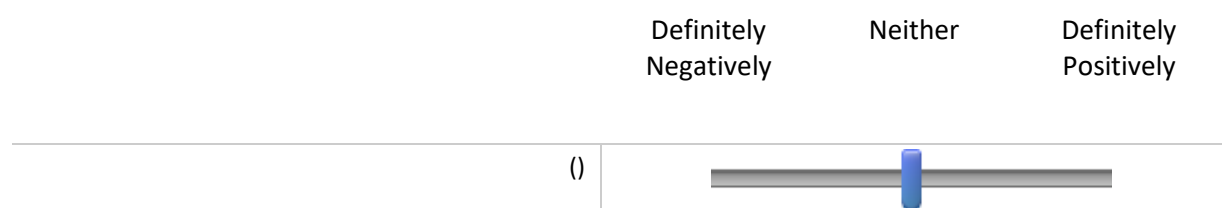


strength Do you think the punisher is...



Page Break

negPos Do you think punishing reflected more negatively or more positively on the punisher?

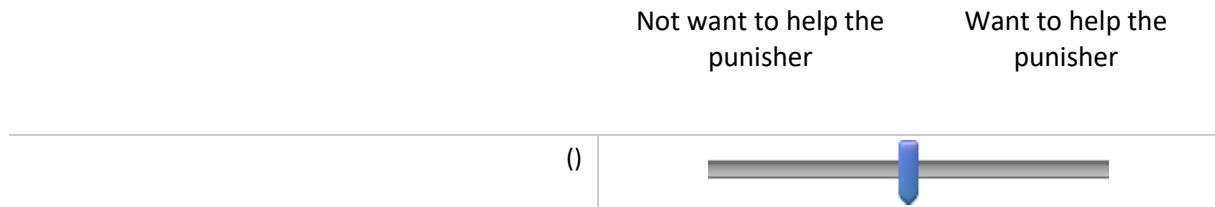


---

Page Break

---

help If the opportunity were to present itself in future, would you:

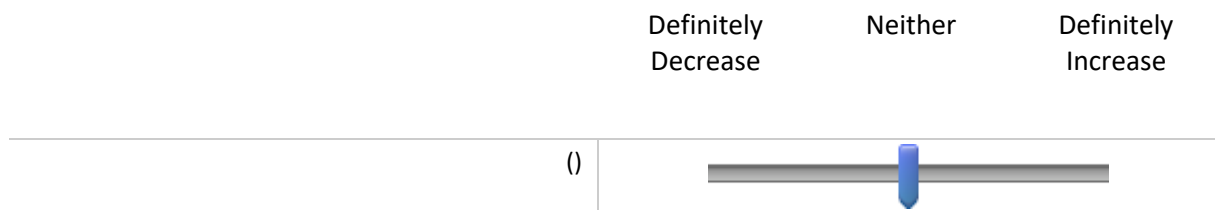


---

Page Break

---

associate Did the punishing behaviour decrease or increase your willingness to associate (e.g., be friends) with the punisher?



End of Block: Punisher Evaluation

---

Start of Block: Alternative



alternative Alternatively, please recall a recent punishing behaviour you have for instance seen on TV (e.g., in a fictional show) or read about (e.g., in a book or magazine).

The people involved in this punishment action do not have to be real, but they must be people (e.g., not country-level punishment, not anthropomorphised animals).

If multiple punishing acts come to mind, please **choose only ONE** to report on throughout this survey.

source Please briefly tell us what the source of the punishing behaviour you are reporting on is (e.g., "fantasy film" or "reality TV").

---

describeA Please briefly describe the *punishing act*.

**What action(s) did the PUNISHER take in response to the wrong behaviour?**

Although it can be useful to mention what violation was punished (i.e., what the other person did to be punished), we are interested in the *punishing behaviour* (i.e., *how the violator was punished*).

---

---

---

Page Break



formA What form did the punishment take?

*If you select "Other", please expand.*

- ☐ Material (1)
- ☐ Avoidance (2)
- ☐ Exclusion (3)
- ☐ Withdrawal of benefits (4)
- ☐ Gossip (5)
- ☐ Verbal Confrontation (6)
- ☐ Physical Confrontation (7)
- ☐ Discipline (8)
- ☐ Other (9) \_\_\_\_\_

Page Break

severeA How severe was the violation that was punished?

Not at all severe

Very severe

()



Page Break

justifiedA How justified was the punishment?

Not at all justified

Completely justified

()



Page Break



intentA Do you believe the violation was accidental or intentional?

☐ Accidental (0)

☐ Intentional (1)

Page Break

vicA **Who was the victim of the violation (i.e., who was affected by the violators wrong behaviour)?**  
Select all that apply.

- ☐ The Punisher (2)
- ☐ Someone Else (3)
- ☐ The general community / environment (4)



viola relshpA **What relationship do the *punisher* and the *viola* have to each other?**  
The violator is the person(s) who engaged in the wrong behaviour that was punished.

- ☐ Strangers (1)
- ☐ Acquaintances (2)
- ☐ Colleagues/ Classmates (3)
- ☐ Competitors (4)
- ☐ Friends (5)
- ☐ Romantic Partners (6)
- ☐ Family (7)
- ☐ Other (8) \_\_\_\_\_



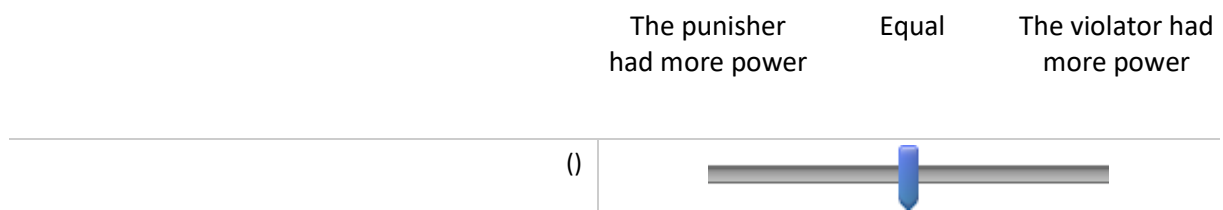
vic relshpA **What relationship do the *punisher* and the *victim* have to each other?**  
The victim is the person(s) affected by the violator's wrong behaviour.

*Note:* If you previously indicated that the punisher is the victim you should select the first option here.

- ☐ The punisher was the victim (1)
- ☐ Strangers (2)
- ☐ Acquaintances (3)
- ☐ Colleagues/ Classmates (4)
- ☐ Competitors (5)
- ☐ Friends (6)
- ☐ Romantic Partners (7)
- ☐ Family (8)
- ☐ Other (9) \_\_\_\_\_

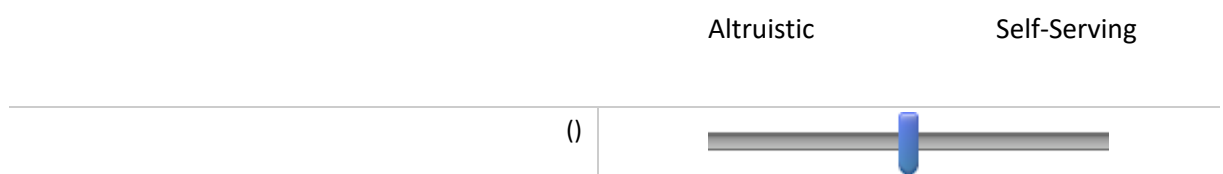
Page Break

powerA Was there any power asymmetry between the punisher and the violator? This includes higher social status, dominance, prestige etc.



Page Break

motiveA Do you think that the punishers motivation for punishing was more altruistic or more self-serving?



Page Break

likeA Do you think the punisher is...

Not at all likeable

Very likeable

()



TrustA Do you think the punisher is...

Very Untrustworthy

Very Trustworthy

()



generousA Do you think the punisher is...

Not at all generous

Very generous

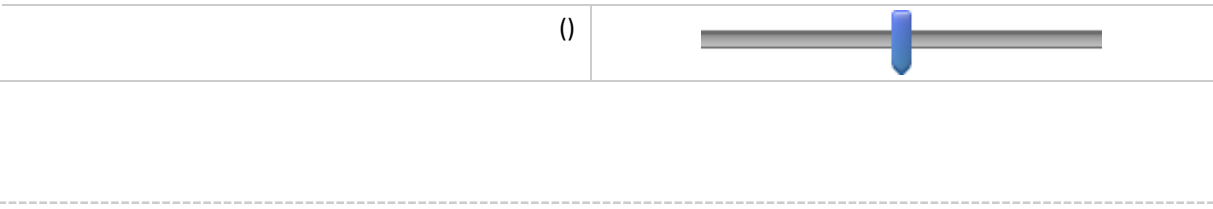
()



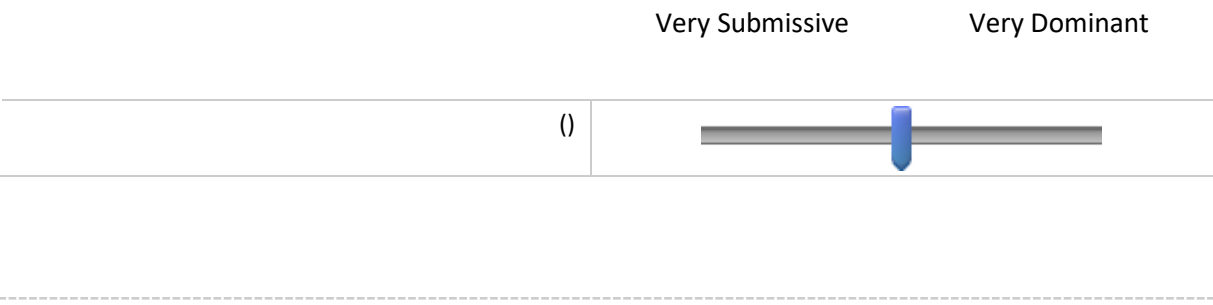
coopA Do you think the punisher is...

Very Uncooperative

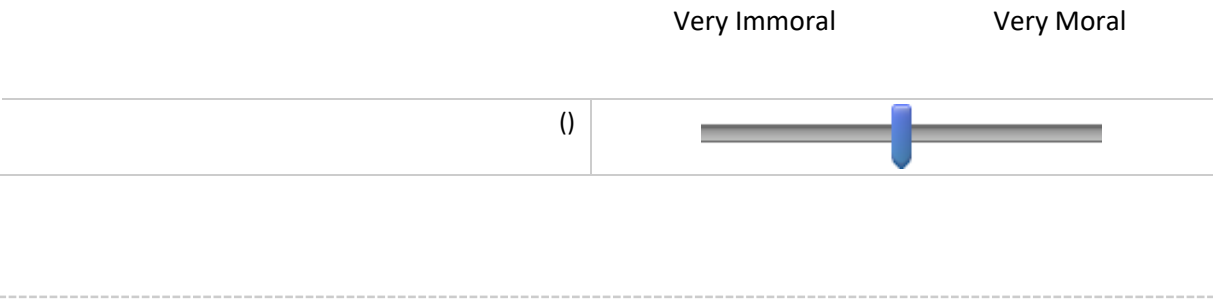
Very Cooperative



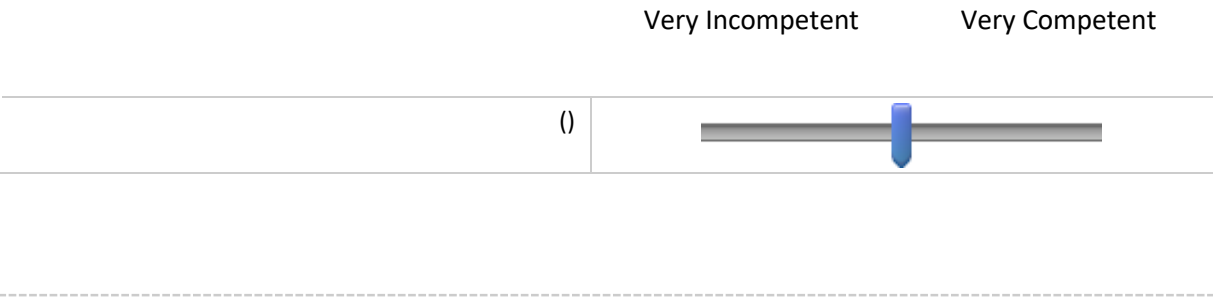
subDomA Do you think the punisher is...



moralA Do you think the punisher is...

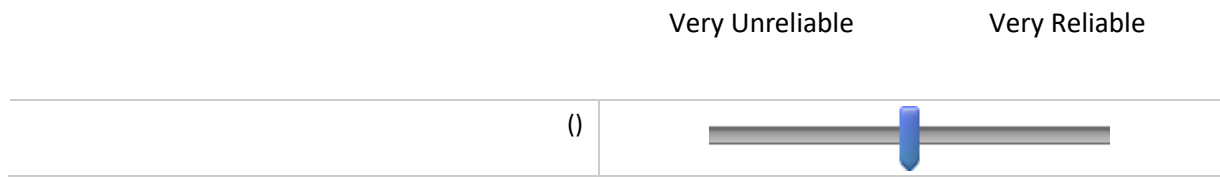


competentA Do you think the punisher is...





reliableA Do you think the punisher is...

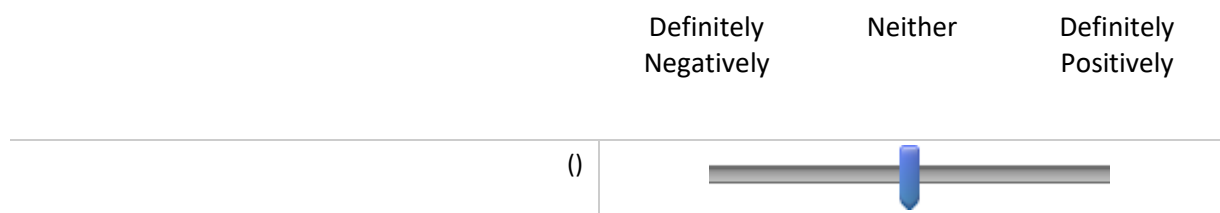


strengthA Do you think the punisher is...



Page Break

posNegA Do you think punishing reflected more negatively or more positively on the punisher?



Page Break

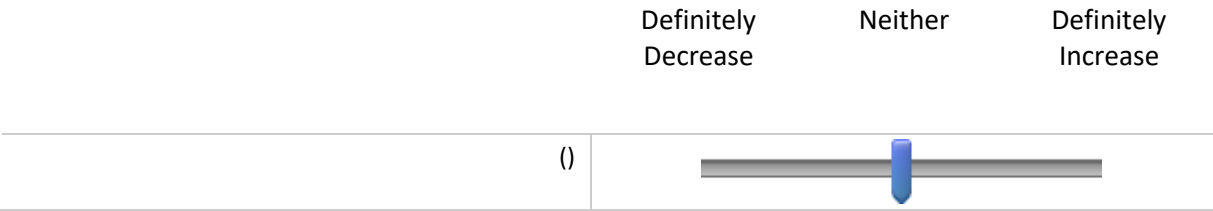
helpA If the punisher needed help with something and you were able to help, would you:

Not want to help the punisher                      Want to help the punisher



Page Break

assocA Would the punishing behaviour have decreased or increased your willingness to associate (e.g., be friends) with the punisher?



End of Block: Alternative

Start of Block: comment

comments Do you have any comments? Optional.

redirect Please click the button below to be redirected back to Prolific and record your submission.

End of Block: comment

## **B2   Analyses Including Non-Real Life Punishing Instances**

**Table B2**

*The association between context variables and punishers' reputations in real life and non-real-life situations*

|                          | <b>Punishment<br/>Reflection</b>    | <b>Trust-<br/>worthiness</b>        | <b>Generosity</b>                   | <b>Reliability</b>                  | <b>Cooperative<br/>-ness</b>        | <b>Morality</b>                     | <b>Likeability</b>                  | <b>Competence</b>                   | <b>Toughness</b>                    | <b>Dominance</b>                    | <b>Willingness<br/>to help</b>      | <b>Willingness<br/>to associate</b> |
|--------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| <i>Predictors</i>        | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    |
| (Intercept)              | <b>23.20 ***</b><br>(17.72 – 28.68) | <b>42.31 ***</b><br>(36.35 – 48.27) | <b>29.97 ***</b><br>(23.20 – 36.74) | <b>47.82 ***</b><br>(41.76 – 53.88) | <b>36.80 ***</b><br>(30.26 – 43.35) | <b>42.85 ***</b><br>(37.00 – 48.71) | <b>32.92 ***</b><br>(26.38 – 39.47) | <b>53.29 ***</b><br>(47.94 – 58.65) | <b>60.10 ***</b><br>(54.52 – 65.69) | <b>81.58 ***</b><br>(75.54 – 87.61) | <b>20.44 ***</b><br>(13.73 – 27.14) | <b>21.80 ***</b><br>(16.12 – 27.48) |
| Intentional<br>Violation | 3.53<br>(0.49 – 6.57)               |                                     | <b>4.56 **</b><br>(1.27 – 7.85)     |                                     | <b>4.02 *</b><br>(0.82 – 7.23)      | 1.09<br>(-1.81 – 3.99)              | <b>3.55 *</b><br>(0.31 – 6.80)      | 2.24<br>(-0.76 – 5.23)              |                                     | -0.45<br>(-3.54 – 2.65)             | 2.95<br>(-0.27 – 6.17)              | <b>3.79 **</b><br>(0.92 – 6.65)     |
| Justified<br>Punishment  | <b>0.56 ***</b><br>(0.51 – 0.62)    | <b>0.45 ***</b><br>(0.40 – 0.50)    | <b>0.39 ***</b><br>(0.33 – 0.45)    | <b>0.38 ***</b><br>(0.33 – 0.43)    | <b>0.45 ***</b><br>(0.39 – 0.50)    | <b>0.47 ***</b><br>(0.42 – 0.52)    | <b>0.49 ***</b><br>(0.44 – 0.55)    | <b>0.34 ***</b><br>(0.29 – 0.39)    | <b>0.09 **</b><br>(0.04 – 0.14)     | <b>-0.09 **</b><br>(-0.15 – -0.04)  | <b>0.59 ***</b><br>(0.54 – 0.65)    | <b>0.45 ***</b><br>(0.40 – 0.50)    |
| Punishing<br>Motive      | <b>-0.22 ***</b><br>(-0.27 – -0.17) | <b>-0.24 ***</b><br>(-0.29 – -0.19) | <b>-0.24 ***</b><br>(-0.29 – -0.18) | <b>-0.15 ***</b><br>(-0.20 – -0.10) | <b>-0.26 ***</b><br>(-0.31 – -0.21) | <b>-0.19 ***</b><br>(-0.24 – -0.14) | <b>-0.22 ***</b><br>(-0.27 – -0.17) | <b>-0.14 ***</b><br>(-0.19 – -0.09) | -0.02<br>(-0.07 – 0.03)             | 0.03<br>(-0.02 – 0.08)              | <b>-0.19 ***</b><br>(-0.24 – -0.14) | <b>-0.15 ***</b><br>(-0.19 – -0.10) |
| Violation<br>Severity    | 0.04<br>(-0.01 – 0.08)              | -0.01<br>(-0.06 – 0.04)             | -0.04<br>(-0.09 – 0.02)             | -0.02<br>(-0.07 – 0.02)             | -0.05<br>(-0.10 – 0.00)             | -0.04<br>(-0.08 – 0.01)             | -0.03<br>(-0.08 – 0.02)             | 0.01<br>(-0.04 – 0.05)              | <b>0.14 ***</b><br>(0.09 – 0.19)    | <b>0.09 **</b><br>(0.04 – 0.14)     | 0.01<br>(-0.04 – 0.06)              | -0.01<br>(-0.05 – 0.04)             |
| Power<br>Dynamics        | 0.03<br>(-0.02 – 0.08)              |                                     | 0.04<br>(-0.01 – 0.10)              | -0.02<br>(-0.08 – 0.03)             | -0.03<br>(-0.08 – 0.03)             |                                     | 0.03<br>(-0.03 – 0.09)              | -0.03<br>(-0.08 – 0.02)             | <b>-0.10 **</b><br>(-0.16 – -0.05)  | <b>-0.21 ***</b><br>(-0.26 – -0.15) | 0.06<br>(-0.00 – 0.11)              | 0.06<br>(0.01 – 0.11)               |
| Mixed<br>Punishment      |                                     | 4.28<br>(0.20 – 8.36)               | 3.12<br>(-1.20 – 7.43)              | 3.22<br>(-0.85 – 7.30)              | 2.30<br>(-1.89 – 6.48)              |                                     | 2.55<br>(-1.72 – 6.82)              | 1.95<br>(-1.99 – 5.89)              |                                     | -3.22<br>(-7.29 – 0.86)             |                                     | 3.53<br>(-0.27 – 7.34)              |
| SPP                      |                                     | 4.26<br>(1.12 – 7.40)               | <b>5.39 *</b><br>(2.04 – 8.73)      | 1.53<br>(-1.61 – 4.67)              | 3.79<br>(0.55 – 7.03)               |                                     | 2.86<br>(-0.45 – 6.17)              | 3.73<br>(0.82 – 6.64)               |                                     | -1.62<br>(-4.76 – 1.53)             |                                     | 0.77<br>(-2.04 – 3.59)              |

|                                                    | Punishment<br>Reflection | Trust-<br>worthiness                                | Generosity                                           | Reliability                           | Cooperative-<br>ness                                | Morality                                            | Likeability                                         | Competence        | Toughness                                            | Dominance                                           | Willingness<br>to help                              | Willingness<br>to associate |
|----------------------------------------------------|--------------------------|-----------------------------------------------------|------------------------------------------------------|---------------------------------------|-----------------------------------------------------|-----------------------------------------------------|-----------------------------------------------------|-------------------|------------------------------------------------------|-----------------------------------------------------|-----------------------------------------------------|-----------------------------|
| <i>Predictors</i>                                  | <i>Estimates</i>         | <i>Estimates</i>                                    | <i>Estimates</i>                                     | <i>Estimates</i>                      | <i>Estimates</i>                                    | <i>Estimates</i>                                    | <i>Estimates</i>                                    | <i>Estimates</i>  | <i>Estimates</i>                                     | <i>Estimates</i>                                    | <i>Estimates</i>                                    | <i>Estimates</i>            |
| Violator<br>Relationship<br>[Acquaintanc<br>es]    |                          | 1.32<br>(-<br>3.68 – 6.33)                          | 6.24<br>(0.93 – 11.55<br>)                           | -1.58<br>(-<br>6.59 – 3.43)           | 2.33<br>(-<br>2.82 – 7.47)                          | -3.02<br>(-<br>7.74 – 1.70)                         | 2.28<br>(-<br>2.97 – 7.52)                          |                   | -2.69<br>(-<br>7.68 – 2.31)                          | 2.23<br>(-<br>2.79 – 7.26)                          | 3.15<br>(-<br>2.14 – 8.45)                          | 3.23<br>(-<br>1.48 – 7.94)  |
| Violator<br>Relationship<br>[Colleagues]           |                          | 3.57<br>(-<br>0.68 – 7.81)                          | 6.69<br>(2.16 – 11.23<br>)                           | 4.57<br>(0.32 – 8.82)                 | 5.91<br>(1.53 – 10.29<br>)                          | 1.84<br>(-<br>2.17 – 5.85)                          | 3.28<br>(-<br>1.18 – 7.75)                          |                   | -4.14<br>(-<br>8.38 – 0.11)                          | -3.77<br>(-<br>8.05 – 0.50)                         | 1.96<br>(-<br>2.55 – 6.47)                          | 1.52<br>(-<br>2.49 – 5.52)  |
| Violator<br>Relationship<br>[Competitors]          |                          | -0.98<br>(-<br>11.17 – 9.21)                        | 6.47<br>(-<br>4.43 – 17.37)                          | 10.81<br>(0.61 – 21.02<br>)           | -1.59<br>(-<br>12.19 – 9.00)                        | -1.74<br>(-<br>11.43 – 7.95)                        | 5.00<br>(-<br>5.75 – 15.74)                         |                   | 6.95<br>(-<br>3.29 – 17.18)                          | 2.76<br>(-<br>7.56 – 13.07)                         | 2.10<br>(-<br>8.68 – 12.88)                         | 1.49<br>(-<br>8.06 – 11.03) |
| Violator<br>Relationship<br>[Family]               |                          | <b>8.24 ***</b><br><b>(4.55 – 11.94</b><br><b>)</b> | <b>13.95 ***</b><br><b>(9.91 – 17.99</b><br><b>)</b> | <b>5.87 *</b><br><b>(2.19 – 9.56)</b> | <b>9.22 ***</b><br><b>(5.28 – 13.15</b><br><b>)</b> | <b>7.70 ***</b><br><b>(4.19 – 11.21</b><br><b>)</b> | <b>9.87 ***</b><br><b>(5.92 – 13.83</b><br><b>)</b> |                   | <b>-6.25 *</b><br><b>(-10.00 – -</b><br><b>2.51)</b> | <b>-5.78 *</b><br><b>(-9.58 – -</b><br><b>1.97)</b> | <b>7.85 ***</b><br><b>(3.90 – 11.80</b><br><b>)</b> | 4.27<br>(0.81 – 7.73)       |
| Violator<br>Relationship<br>[Friends]              |                          | 5.96<br>(0.47 – 11.44<br>)                          | <b>12.56 ***</b><br><b>(6.75 – 18.37</b><br><b>)</b> | 3.72<br>(-<br>1.75 – 9.20)            | 4.37<br>(-<br>1.26 – 10.00)                         | 5.68<br>(0.55 – 10.81<br>)                          | <b>9.21 *</b><br><b>(3.46 – 14.95</b><br><b>)</b>   |                   | -6.66<br>(-12.09 – -<br>1.23)                        | -5.85<br>(-11.34 – -<br>0.36)                       | 7.57<br>(1.80 – 13.34<br>)                          | 6.29<br>(1.17 – 11.41<br>)  |
| Violator<br>Relationship<br>[Other]                |                          | 5.57 *<br>(1.00 – 10.13<br>)                        | 2.90<br>(-<br>2.04 – 7.84)                           | 3.46<br>(-<br>1.14 – 8.05)            | -0.36<br>(-<br>5.09 – 4.37)                         | <b>4.64 *</b><br><b>(0.32 – 8.95)</b>               | 2.71<br>(-<br>2.13 – 7.55)                          |                   | 1.39<br>(-<br>3.24 – 6.03)                           | 2.27<br>(-<br>2.39 – 6.93)                          | <b>5.66 *</b><br><b>(0.71 – 10.60</b><br><b>)</b>   | 2.46<br>(-<br>1.91 – 6.83)  |
| Violator<br>Relationship<br>[Romantic<br>Partners] |                          | 7.60<br>(1.90 – 13.30<br>)                          | <b>14.80 ***</b><br><b>(8.75 – 20.84</b><br><b>)</b> | 2.59<br>(-<br>2.96 – 8.14)            | <b>9.51 *</b><br><b>(3.52 – 15.50</b><br><b>)</b>   | <b>8.73 **</b><br><b>(3.58 – 13.88</b><br><b>)</b>  | <b>11.06 **</b><br><b>(5.08 – 17.04</b><br><b>)</b> |                   | -4.12<br>(-<br>9.51 – 1.27)                          | -8.07<br>(-13.63 – -<br>2.50)                       | <b>10.95 **</b><br><b>(5.17 – 16.73</b><br><b>)</b> | 7.02<br>(1.88 – 12.16<br>)  |
| N                                                  | 206 <sub>ID</sub>        | 206 <sub>ID</sub>                                   | 206 <sub>ID</sub>                                    | 206 <sub>ID</sub>                     | 206 <sub>ID</sub>                                   | 206 <sub>ID</sub>                                   | 206 <sub>ID</sub>                                   | 206 <sub>ID</sub> | 206 <sub>ID</sub>                                    | 206 <sub>ID</sub>                                   | 206 <sub>ID</sub>                                   | 206 <sub>ID</sub>           |
| Observations                                       | 858                      | 858                                                 | 858                                                  | 858                                   | 858                                                 | 858                                                 | 858                                                 | 858               | 858                                                  | 858                                                 | 857                                                 | 857                         |

\*  $p < 0.0038$  \*\*  $p < 0.001$  \*\*\*  $p < 0.0001$

*Note.* Missing cells indicate that the specific factor was not included in the final top model. Estimates reported are conditional model averages, which average over the models where the parameters appear. The reference level for Violator Relationship is “Strangers”, “TPP” for Punishing Type, and “Accidental” for Violation Intentionality. Significant results are bolded. Estimates are followed by 95% CI in brackets

## B3 Global model results without model selection or model averaging

**Table B3**

*The influence of predictors across a punisher's reputation*

|                          | Punishment<br>Reflection        | Trustworthin<br>ess             | Generosity                      | Reliability                     | Cooperativen<br>ess             | Morality                        | Likeability                     | Competence                      | Toughness                       | Dominance                       | Willingnes<br>s to help         | Willingness<br>to associate     |
|--------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| Predictors               | Estimates                       | Estimates                       | Estimates                       | Estimates                       | Estimates                       | Estimates                       | Estimates                       | Estimates                       | Estimates                       | Estimates                       | Estimates                       | Estimates                       |
| (Intercept)              | <b>19.49</b><br>(12.20 – 26.77) | <b>40.87</b><br>(33.83 – 47.91) | 25.55<br>(17.71 – 33.40)        | <b>45.76</b><br>(38.51 – 53.00) | <b>33.31</b><br>(25.73 – 40.88) | <b>39.87</b><br>(33.09 – 46.65) | <b>30.58</b><br>(23.15 – 38.01) | <b>49.42</b><br>(42.25 – 56.60) | <b>65.02</b><br>(57.64 – 72.41) | <b>80.10</b><br>(72.59 – 87.60) | <b>15.71</b><br>(7.71 – 23.71)  | <b>19.67</b><br>(12.71 – 26.64) |
| Punishing<br>Motive      | <b>-0.19</b><br>(-0.25 – -0.13) | <b>-0.20</b><br>(-0.26 – -0.14) | <b>-0.18</b><br>(-0.24 – -0.12) | <b>-0.13</b><br>(-0.19 – -0.07) | <b>-0.21</b><br>(-0.27 – -0.15) | <b>-0.15</b><br>(-0.20 – -0.09) | <b>-0.19</b><br>(-0.25 – -0.13) | <b>-0.14</b><br>(-0.20 – -0.09) | -0.05<br>(-0.11 – 0.01)         | 0.03<br>(-0.03 – 0.09)          | <b>-0.16</b><br>(-0.22 – -0.10) | <b>-0.12</b><br>(-0.18 – -0.07) |
| Mixed<br>Punishment      | -1.09<br>(-5.51 – 3.34)         | 5.31<br>(0.96 – 9.66)           | 3.09<br>(-1.72 – 7.90)          | 3.90<br>(-0.55 – 8.34)          | 2.09<br>(-2.55 – 6.72)          | 2.22<br>(-1.85 – 6.30)          | 2.66<br>(-1.91 – 7.23)          | 1.45<br>(-2.93 – 5.83)          | 0.20<br>(-4.31 – 4.70)          | -1.26<br>(-5.94 – 3.42)         | 0.16<br>(-4.71 – 5.03)          | 2.95<br>(-1.27 – 7.17)          |
| SPP                      | -0.93<br>(-4.35 – 2.49)         | 2.94<br>(-0.43 – 6.31)          | 3.32<br>(-0.40 – 7.05)          | 1.78<br>(-1.66 – 5.23)          | 1.30<br>(-2.29 – 4.88)          | 1.60<br>(-1.55 – 4.75)          | 1.55<br>(-1.99 – 5.09)          | 3.69<br>(0.30 – 7.08)           | -1.23<br>(-4.72 – 2.26)         | -1.11<br>(-4.74 – 2.52)         | -0.36<br>(-4.14 – 3.42)         | -2.98<br>(-6.25 – 0.29)         |
| Justified<br>Punishment  | <b>0.59</b><br>(0.52 – 0.65)    | <b>0.46</b><br>(0.39 – 0.52)    | <b>0.39</b><br>(0.33 – 0.46)    | <b>0.38</b><br>(0.32 – 0.44)    | <b>0.48</b><br>(0.42 – 0.55)    | <b>0.43</b><br>(0.37 – 0.49)    | <b>0.50</b><br>(0.44 – 0.57)    | <b>0.36</b><br>(0.29 – 0.42)    | 0.06<br>(-0.01 – 0.12)          | -0.07<br>(-0.14 – -0.01)        | <b>0.58</b><br>(0.51 – 0.65)    | <b>0.42</b><br>(0.36 – 0.48)    |
| Intentional<br>Violation | 2.72<br>(-0.74 – 6.17)          | -0.23<br>(-3.64 – 3.18)         | 4.70<br>(0.94 – 8.46)           | 0.72<br>(-2.76 – 4.20)          | 4.06<br>(0.44 – 7.68)           | 1.90<br>(-1.28 – 5.07)          | 4.53<br>(0.95 – 8.11)           | 1.31<br>(-2.11 – 4.73)          | -0.29<br>(-3.81 – 3.23)         | -0.16<br>(-3.83 – 3.52)         | 3.63<br>(-0.18 – 7.43)          | 3.48<br>(0.18 – 6.77)           |
| Violation<br>Severity    | 0.03<br>(-0.03 – 0.09)          | 0.01<br>(-0.04 – 0.07)          | 0.03<br>(-0.04 – 0.09)          | -0.00<br>(-0.06 – 0.06)         | -0.02<br>(-0.08 – 0.04)         | 0.01<br>(-0.04 – 0.07)          | -0.02<br>(-0.08 – 0.04)         | 0.01<br>(-0.04 – 0.07)          | <b>0.11</b><br>(0.05 – 0.17)    | 0.05<br>(-0.01 – 0.11)          | 0.01<br>(-0.05 – 0.08)          | 0.03<br>(-0.03 – 0.08)          |
| Power<br>Dynamics        | 0.02<br>(-0.04 – 0.09)          | -0.04<br>(-0.10 – 0.02)         | 0.03<br>(-0.03 – 0.10)          | -0.05<br>(-0.11 – 0.01)         | -0.04<br>(-0.10 – 0.03)         | 0.01<br>(-0.05 – 0.06)          | 0.01<br>(-0.06 – 0.07)          | -0.03<br>(-0.09 – 0.03)         | <b>-0.12</b><br>(-0.18 – -0.06) | <b>-0.20</b><br>(-0.26 – -0.13) | 0.07<br>(-0.00 – 0.13)          | 0.07<br>(0.01 – 0.13)           |

|                                                      |                          |                                      |                                        |                                      |                                      |                                      |                                       |                            |                                  |                                  |                                       |                          |
|------------------------------------------------------|--------------------------|--------------------------------------|----------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|---------------------------------------|----------------------------|----------------------------------|----------------------------------|---------------------------------------|--------------------------|
| Violator Relationship [Acquaintances]                | -2.53<br>(- 8.40 – 3.34) | 4.48<br>(-1.29 – 10.26)              | <b>9.70</b><br><b>(3.31 – 16.08)</b>   | 1.45<br>(- 4.46 – 7.35)              | 3.88<br>(-2.27 – 10.03)              | -1.75<br>(- 7.16 – 3.65)             | 4.85<br>(- 1.22 – 10.92)              | 1.15<br>(- 4.66 – 6.96)    | -0.13<br>(- 6.11 – 5.84)         | 1.58<br>(- 4.64 – 7.79)          | 5.70<br>(- 0.77 – 12.16)              | 3.35<br>(- 2.25 – 8.95)  |
| Violator Relationship [Colleagues/Classmates]        | -1.00<br>(- 5.58 – 3.58) | 3.51<br>(-0.96 – 7.99)               | 6.37<br>(1.41 – 11.34)                 | 6.28<br>(1.70 – 10.87)               | 5.69<br>(0.90 – 10.47)               | 1.32<br>(- 2.91 – 5.55)              | 3.22<br>(- 1.50 – 7.93)               | 4.60<br>(0.07 – 9.12)      | -1.95<br>(- 6.61 – 2.70)         | -3.07<br>(- 7.87 – 1.73)         | 2.97<br>(- 2.07 – 8.01)               | 0.95<br>(- 3.43 – 5.32)  |
| Violator Relationship [Competitors]                  | 5.96<br>(- 9.59 – 21.52) | -6.68<br>(-22.10 – 8.75)             | 3.18<br>(- 13.79 – 20.16)              | -0.62<br>(- 16.32 – 15.08)           | -0.85<br>(- 17.19 – 15.50)           | -7.29<br>(- 21.55 – 6.97)            | -11.49<br>(- 27.67 – 4.68)            | -12.23<br>(- 27.66 – 3.21) | 4.43<br>(- 11.43 – 20.29)        | 6.25<br>(- 10.41 – 22.92)        | 12.06<br>(- 5.09 – 29.20)             | 8.41<br>(- 6.42 – 23.24) |
| Violator Relationship [Family]                       | 2.02<br>(- 2.23 – 6.27)  | <b>9.28</b><br><b>(5.13 – 13.44)</b> | <b>16.41</b><br><b>(11.80 – 21.01)</b> | <b>7.11</b><br><b>(2.85 – 11.36)</b> | <b>9.09</b><br><b>(4.64 – 13.53)</b> | <b>8.57</b><br><b>(4.64 – 12.50)</b> | <b>11.56</b><br><b>(7.19 – 15.94)</b> | 4.62<br>(0.42 – 8.83)      | -6.04<br><b>(-10.37 – -1.72)</b> | -6.24<br><b>(-10.69 – -1.79)</b> | <b>11.31</b><br><b>(6.63 – 15.99)</b> | 5.09<br>(1.03 – 9.16)    |
| Violator Relationship [Friends]                      | 2.32<br>(- 3.72 – 8.36)  | 5.59<br>(-0.37 – 11.55)              | <b>13.68</b><br><b>(7.10 – 20.25)</b>  | 3.77<br>(- 2.31 – 9.85)              | 4.86<br>(-1.48 – 11.19)              | 5.32<br>(- 0.24 – 10.88)             | <b>9.63</b><br><b>(3.37 – 15.88)</b>  | 3.23<br>(- 2.76 – 9.21)    | -3.89<br>(- 10.05 – 2.26)        | -5.16<br>(- 11.58 – 1.26)        | <b>10.74</b><br><b>(4.08 – 17.39)</b> | 5.26<br>(- 0.50 – 11.03) |
| Violator Relationship [Other]                        | 1.31<br>(- 3.74 – 6.36)  | 4.42<br>(-0.51 – 9.36)               | 2.56<br>(- 2.91 – 8.03)                | 2.56<br>(- 2.49 – 7.62)              | -2.53<br>(-7.81 – 2.74)              | 3.65<br>(- 1.01 – 8.32)              | 1.98<br>(- 3.21 – 7.18)               | 0.12<br>(- 4.87 – 5.11)    | 2.30<br>(- 2.84 – 7.43)          | 2.94<br>(- 2.35 – 8.23)          | 7.36<br>(1.80 – 12.92)                | 3.24<br>(- 1.59 – 8.06)  |
| Violator Relationship [Romantic Partners]            | 1.79<br>(- 5.77 – 9.36)  | 7.64<br>(0.22 – 15.06)               | <b>18.51</b><br><b>(10.30 – 26.72)</b> | 2.39<br>(- 5.20 – 9.98)              | 9.34<br>(1.42 – 17.25)               | 3.69<br>(- 3.29 – 10.67)             | <b>12.02</b><br><b>(4.22 – 19.83)</b> | 2.39<br>(- 5.10 – 9.87)    | -3.51<br>(- 11.21 – 4.19)        | -9.96<br>(-17.94 – -1.99)        | <b>13.83</b><br><b>(5.50 – 22.16)</b> | 7.13<br>(- 0.10 – 14.35) |
| N                                                    | 190 <sub>ID</sub>        | 190 <sub>ID</sub>                    | 190 <sub>ID</sub>                      | 190 <sub>ID</sub>                    | 190 <sub>ID</sub>                    | 190 <sub>ID</sub>                    | 190 <sub>ID</sub>                     | 190 <sub>ID</sub>          | 190 <sub>ID</sub>                | 190 <sub>ID</sub>                | 190 <sub>ID</sub>                     | 190 <sub>ID</sub>        |
| Observations                                         | 638                      | 638                                  | 638                                    | 638                                  | 638                                  | 638                                  | 638                                   | 638                        | 638                              | 638                              | 637                                   | 637                      |
| Marginal R <sup>2</sup> / Conditional R <sup>2</sup> | 0.590 / 0.680            | 0.509 / 0.575                        | 0.427 / 0.526                          | 0.383 / 0.487                        | 0.506 / 0.596                        | 0.479 / 0.621                        | 0.523 / 0.596                         | 0.365 / 0.487              | 0.125 / 0.299                    | 0.109 / 0.196                    | 0.527 / 0.624                         | 0.460 / 0.583            |

*Note.* Results from 12 global linear mixed-effects models (one for each of the 12 dependent variables) are presented, showing coefficients and 95% confidence intervals (CIs). Significant results ( $p < .0038$ ) are bolded, and those that deviate from the significance levels observed in the multi-model selection with model averaging method reported in the main text are highlighted in yellow.

## **B4 Analyses Including Interactions Between Motive and Punishing Type**



Table B4.1

*The association between context variables and punishers' reputations when interactions between motivation and punishing type are included*

|                                                 | Punishment<br>Reflection            | Trust-<br>worthiness                | Generosity                          | Reliability                         | Cooperative<br>-ness                | Morality                            | Likeability                         | Competence                          | Toughness                           | Dominance                           | Willingness<br>to help            | Willingness<br>to associate         |
|-------------------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-----------------------------------|-------------------------------------|
| <i>Predictors</i>                               | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                    | <i>Estimates</i>                  | <i>Estimates</i>                    |
| (Intercept)                                     | <b>21.06 ***</b><br>(14.76 – 27.36) | <b>40.28 ***</b><br>(33.46 – 47.10) | <b>26.82 ***</b><br>(19.18 – 34.46) | <b>46.06 ***</b><br>(38.34 – 53.77) | <b>32.35 ***</b><br>(25.04 – 39.66) | <b>41.02 ***</b><br>(34.63 – 47.41) | <b>30.69 ***</b><br>(23.66 – 37.73) | <b>52.05 ***</b><br>(45.80 – 58.30) | <b>64.81 ***</b><br>(56.05 – 73.58) | <b>80.41 ***</b><br>(72.76 – 88.06) | <b>15.56 **</b><br>(7.53 – 23.59) | <b>22.04 ***</b><br>(15.67 – 28.41) |
| Intentional<br>Violation                        | 3.14<br>(-0.19 – 6.48)              |                                     | 4.86<br>(1.15 – 8.57)               | 0.73<br>(-2.64 – 4.10)              | 3.82<br>(0.31 – 7.33)               | 1.95<br>(-1.12 – 5.02)              | 4.39<br>(0.92 – 7.86)               | 1.45<br>(-1.91 – 4.82)              |                                     |                                     | 3.79<br>(0.05 – 7.53)             | 4.08<br>(0.83 – 7.33)               |
| Justified<br>Punishment                         | <b>0.59 ***</b><br>(0.53 – 0.66)    | <b>0.46 ***</b><br>(0.40 – 0.52)    | <b>0.40 ***</b><br>(0.33 – 0.47)    | <b>0.38 ***</b><br>(0.32 – 0.44)    | <b>0.48 ***</b><br>(0.42 – 0.54)    | <b>0.44 ***</b><br>(0.39 – 0.50)    | <b>0.50 ***</b><br>(0.44 – 0.56)    | <b>0.36 ***</b><br>(0.30 – 0.42)    | 0.06<br>(-0.00 – 0.13)              | -0.07<br>(-0.13 – 0.01)             | <b>0.59 ***</b><br>(0.52 – 0.66)  | <b>0.42 ***</b><br>(0.36 – 0.48)    |
| Punishing<br>Motive                             | <b>-0.20 ***</b><br>(-0.25 – -0.14) | <b>-0.20 ***</b><br>(-0.26 – -0.15) | <b>-0.17 ***</b><br>(-0.24 – -0.10) | <b>-0.13 ***</b><br>(-0.18 – -0.07) | <b>-0.20 ***</b><br>(-0.26 – -0.15) | <b>-0.14 ***</b><br>(-0.19 – -0.09) | <b>-0.18 ***</b><br>(-0.24 – -0.12) | <b>-0.14 ***</b><br>(-0.20 – -0.09) | -0.05<br>(-0.13 – 0.03)             | 0.03<br>(-0.05 – 0.10)              | -0.10<br>(-0.18 – 0.02)           | -0.09<br>(-0.17 – 0.02)             |
| Violation<br>Severity                           | 0.02<br>(-0.03 – 0.08)              | 0.01<br>(-0.04 – 0.07)              | 0.03<br>(-0.04 – 0.09)              |                                     | -0.02<br>(-0.08 – 0.04)             | 0.02<br>(-0.04 – 0.07)              | -0.02<br>(-0.08 – 0.04)             |                                     | <b>0.12 **</b><br>(0.06 – 0.17)     | 0.05<br>(-0.01 – 0.11)              |                                   | 0.02<br>(-0.04 – 0.07)              |
| Power<br>Dynamics                               | 0.02<br>(-0.04 – 0.08)              | -0.04<br>(-0.10 – 0.02)             | 0.03<br>(-0.03 – 0.10)              | -0.05<br>(-0.12 – 0.01)             | -0.04<br>(-0.10 – 0.03)             | 0.01<br>(-0.04 – 0.07)              |                                     | -0.05<br>(-0.10 – 0.01)             | <b>-0.12 **</b><br>(-0.18 – -0.06)  | <b>-0.20 ***</b><br>(-0.26 – -0.13) | 0.07<br>(0.00 – 0.14)             | 0.07<br>(0.01 – 0.12)               |
| Mixed<br>Punishment                             |                                     | 5.29<br>(0.96 – 9.62)               | 3.08<br>(-3.11 – 9.26)              | 4.25<br>(-0.25 – 8.76)              |                                     |                                     |                                     | 2.26<br>(-2.09 – 6.61)              | 6.83<br>(-0.46 – 14.13)             | 5.35<br>(-2.28 – 12.97)             | 1.14<br>(-6.73 – 9.01)            | 3.69<br>(-2.67 – 10.05)             |
| SPP                                             |                                     | 2.92<br>(-0.41 – 6.25)              | 5.68<br>(-1.92 – 13.28)             | 1.87<br>(-1.57 – 5.30)              |                                     |                                     |                                     | 4.18<br>(0.95 – 7.41)               | 2.68<br>(-3.76 – 9.12)              | -6.23<br>(-12.93 – 0.48)            | 9.43<br>(2.46 – 16.41)            | 3.21<br>(-4.14 – 10.55)             |
| Violator<br>Relationship<br>[Acquaintanc<br>es] |                                     | 4.42<br>(-1.34 – 10.19)             | <b>9.86 *</b><br>(3.47 – 16.25)     | 1.46<br>(-4.44 – 7.35)              | 3.82<br>(-2.31 – 9.95)              | -1.54<br>(-6.93 – 3.85)             | 4.93<br>(-1.12 – 10.99)             |                                     | -0.41<br>(-6.39 – 5.57)             | 1.79<br>(-4.42 – 8.00)              | 5.22<br>(-1.20 – 11.65)           |                                     |

|                                                    | Punishment<br>Reflection | Trust-<br>worthiness                                | Generosity                                            | Reliability                                        | Cooperative<br>-ness                                | Morality                                            | Likeability                                          | Competence        | Toughness                                            | Dominance                        | Willingness<br>to help                              | Willingness<br>to associate  |
|----------------------------------------------------|--------------------------|-----------------------------------------------------|-------------------------------------------------------|----------------------------------------------------|-----------------------------------------------------|-----------------------------------------------------|------------------------------------------------------|-------------------|------------------------------------------------------|----------------------------------|-----------------------------------------------------|------------------------------|
| <i>Predictors</i>                                  | <i>Estimates</i>         | <i>Estimates</i>                                    | <i>Estimates</i>                                      | <i>Estimates</i>                                   | <i>Estimates</i>                                    | <i>Estimates</i>                                    | <i>Estimates</i>                                     | <i>Estimates</i>  | <i>Estimates</i>                                     | <i>Estimates</i>                 | <i>Estimates</i>                                    | <i>Estimates</i>             |
| Violator<br>Relationship<br>[Competitors]          |                          | -6.43<br>(-<br>21.74 – 8.87)                        | 4.11<br>(-<br>12.77 – 21.0<br>0)                      | -0.58<br>(-<br>16.13 – 14.9<br>7)                  | -1.32<br>(-<br>17.52 – 14.8<br>9)                   | -6.59<br>(-<br>20.72 – 7.54)                        | -11.75<br>(-<br>27.80 – 4.31)                        |                   | 4.23<br>(-<br>11.63 – 20.0<br>9)                     | 6.20<br>(-<br>10.46 – 22.8<br>7) | 12.76<br>(-<br>4.16 – 29.68<br>)                    |                              |
| Violator<br>Relationship<br>[Family]               |                          | <b>9.54 ***</b><br><b>(5.46 – 13.61</b><br><b>)</b> | <b>15.90 ***</b><br><b>(11.28 – 20.5</b><br><b>2)</b> | <b>7.84 **</b><br><b>(3.66 – 12.03</b><br><b>)</b> | <b>9.99 ***</b><br><b>(5.73 – 14.25</b><br><b>)</b> | <b>8.83 ***</b><br><b>(5.11 – 12.55</b><br><b>)</b> | <b>12.04 ***</b><br><b>(7.91 – 16.18</b><br><b>)</b> |                   | <b>-6.36 *</b><br><b>(-10.62 – -</b><br><b>2.10)</b> | -6.29<br>(-10.78 – -<br>1.81)    | <b>9.74 ***</b><br><b>(5.02 – 14.4</b><br><b>5)</b> |                              |
| Violator<br>Relationship<br>[Friends]              |                          | 5.41<br>(-<br>0.53 – 11.36)                         | <b>14.12 ***</b><br><b>(7.52 – 20.72</b><br><b>)</b>  | 3.96<br>(-<br>2.11 – 10.03)                        | 5.04<br>(-<br>1.24 – 11.31)                         | 5.75<br>(0.25 – 11.25<br>)                          | <b>10.10 *</b><br><b>(3.91 – 16.29</b><br><b>)</b>   |                   | -4.30<br>(-<br>10.41 – 1.82)                         | -5.47<br>(-<br>11.86 – 0.92)     | <b>10.71 *</b><br><b>(4.08 – 17.3</b><br><b>4)</b>  |                              |
| Violator<br>Relationship<br>[Other]                |                          | 4.94<br>(0.04 – 9.85)                               | 2.11<br>(-<br>3.30 – 7.53)                            | 2.99<br>(-<br>2.08 – 8.05)                         | -2.02<br>(-<br>7.23 – 3.19)                         | 3.57<br>(-<br>0.98 – 8.12)                          | 1.88<br>(-<br>3.16 – 6.92)                           |                   | 2.23<br>(-<br>2.90 – 7.36)                           | 3.22<br>(-<br>2.08 – 8.53)       | 6.30<br>(0.70 – 11.9<br>0)                          |                              |
| Violator<br>Relationship<br>[Romantic<br>Partners] |                          | 7.50<br>(0.09 – 14.91<br>)                          | <b>19.44 ***</b><br><b>(11.19 – 27.7</b><br><b>0)</b> | 2.74<br>(-<br>4.76 – 10.23)                        | 9.69<br>(1.99 – 17.39<br>)                          | 4.49<br>(-<br>2.30 – 11.28)                         | <b>12.73 **</b><br><b>(5.15 – 20.31</b><br><b>)</b>  |                   | -4.11<br>(-<br>11.71 – 3.50)                         | -10.02<br>(-17.84 – -<br>2.20)   | <b>14.31 **</b><br><b>(6.03 – 22.5</b><br><b>9)</b> |                              |
| Motive:Mixed<br>Punishment                         |                          |                                                     | -0.00<br>(-<br>0.17 – 0.16)                           |                                                    |                                                     |                                                     |                                                      |                   | -0.18<br>(-0.33 – -<br>0.02)                         | -0.18<br>(-0.34 – -<br>0.02)     | -0.02<br>(-<br>0.19 – 0.14)                         | -0.01<br>(-<br>0.16 – 0.13)  |
| Motive:SPP                                         |                          |                                                     | -0.11<br>(-<br>0.23 – 0.00)                           |                                                    |                                                     |                                                     |                                                      |                   | -0.08<br>(-<br>0.19 – 0.02)                          | 0.09<br>(-<br>0.02 – 0.20)       | <b>-0.19 *</b><br><b>(-0.31 – -</b><br><b>0.08)</b> | -0.12<br>(-0.22 – -<br>0.02) |
| N                                                  | 190 <sub>ID</sub>        | 190 <sub>ID</sub>                                   | 190 <sub>ID</sub>                                     | 190 <sub>ID</sub>                                  | 190 <sub>ID</sub>                                   | 190 <sub>ID</sub>                                   | 190 <sub>ID</sub>                                    | 190 <sub>ID</sub> | 190 <sub>ID</sub>                                    | 190 <sub>ID</sub>                | 190 <sub>ID</sub>                                   | 190 <sub>ID</sub>            |
| Observations                                       | 638                      | 638                                                 | 638                                                   | 638                                                | 638                                                 | 638                                                 | 638                                                  | 638               | 638                                                  | 638                              | 637                                                 | 637                          |

\*  $p < 0.0038$  \*\*  $p < 0.001$  \*\*\*  $p < 0.0001$

*Note.* Missing cells indicate that the specific factor was not included in the final top model. Estimates reported are conditional model averages, which average over the models where the parameters appear. The reference level for Violator Relationship is “Strangers”, “TPP” for Punishing Type, and “Accidental” for Violation Intentionality. Significant results are bolded. Estimates are followed by 95% CI in brackets.

**Table B4.2***Importance of punishing context factors in different models*

|                    | Justified<br>Punishment | Punishing<br>Motive | Violator<br>Relationship | Violation<br>Intentionality | Power<br>Dynamics | Punishing<br>Type | Violation<br>Severity | Motive :<br>Type |
|--------------------|-------------------------|---------------------|--------------------------|-----------------------------|-------------------|-------------------|-----------------------|------------------|
| Overall reflection | 1                       | 1                   | -                        | 0.81                        | 0.18              | -                 | 0.22                  | -                |
| Trustworthiness    | 1                       | 1                   | 1                        | -                           | 0.4               | 1                 | 0.17                  | -                |
| Generosity         | 1                       | 1                   | 1                        | 1                           | 0.30              | 0.60              | 0.24                  | 0.24             |
| Reliability        | 1                       | 1                   | 0.87                     | 0.09                        | 0.64              | 0.44              | -                     | -                |
| Cooperativeness    | 1                       | 1                   | 1                        | 1                           | 0.32              | -                 | 0.2                   | -                |
| Morality           | 1                       | 1                   | 1                        | 0.3                         | 0.15              | -                 | 0.17                  | -                |
| Likeability        | 1                       | 1                   | 1                        | 1                           | -                 | -                 | 0.3                   | -                |
| Competence         | 1                       | 1                   | -                        | 0.34                        | 0.55              | 1                 | -                     | -                |
| Toughness          | 0.77                    | 0.80                | 1                        | -                           | 1                 | 0.18              | 1                     | 0.18             |
| Dominance          | 0.89                    | 0.62                | 1                        | -                           | 1                 | 0.52              | 0.56                  | 0.52             |
| Help willingness   | 1                       | 1                   | 1                        | 0.77                        | 0.77              | 1                 | 0.21                  | 1                |
| Association        | 1                       | 1                   | -                        | 1                           | 1                 | 1                 | 0.24                  | 0.79             |

*Note.* This table shows the importance of each of the variables across different models, reflecting their influence on the twelve response variables. The values indicate the aggregated impact of each variable derived from the `sw()` function in the MuMIn package, which calculates the per-variable sum of model weights. Each value represents how strongly the variable contributes to the models' predictions regarding the response variables listed on the left. A value of 1 means the variable is present in all models and has a strong influence. Values less than 1 indicate varying degrees of importance, with higher values representing greater influence. A dash (-) means the variable was not included in the top models for that response variable.

## B5 A Sample of Free Text Responses

Within this section, a random assortment of 20 participant responses is showcased to provide a brief overview of the diverse range of punishing behaviours reported on by participants. The responses varied in the severity of violations and punishments, the relationships between those involved, the punishing tactics used, the participants' levels of involvement, the detail of descriptions, and the explicitness of judgments. They include both second-party and third-party punishing events, cases involving individuals and groups, and actions occurring both online and in person.

In some cases, punishments involved close, personal relationships, such as family members (*"My cousin who is 15 and lives with her parents took money from her mum's purse and then went out with her friends all night, the next day her parents punished her by grounding her, taking away her phone and all devices and making her do extra chores around the house for a month to pay back the money that she stole"*) and romantic partners (*"My mother refused to make dinner for my father and didn't speak to him because he insulted her in front of everyone"*), while others concerned acquaintances (*"A neighbour had (again) parked his car on double yellow lines. This time he earned himself a parking ticket"*) or even strangers (*"While swimming I inadvertently touched the foot of a lady swimming more slowly in front of me. I think she attempted to punish me for the rest of my swim by deliberately swimming abnormally slowly/pushing off just in front of me during the rest of my swim"*).

The punishments also differed in tactic, spanning from physical confrontation (*"My friend punched a guy in the face because he said rude things about his girlfriend"*) to indirect actions like gossip (*"The punishing act was gossip, which in my mind seems to be the most common type of punishment as it usually avoids direct confrontation. The punisher, SP, in response to a violation, slandered the violator. Firstly, by describing the violation, then relating the violation to numerous bad things in the violators personality"*) or social exclusion (*"Two old ladies had an argument about a third neighbour. As a result of this, old lady no 1 did not ask old lady no 2 to go with her on her next visit to the supermarket. Typically, this is something they would do together every week"*). Those responses also provide examples of variation in description detail and insight.

Some participants described incidents involving second-party punishment (*"My friend charged her sister for a suitcase she lent her that was broken by the airlines when her sister was travelling"*), while other examples highlighted third-party interventions, both online

*(“Facebook community standards banned a friend online for what they considered hate speech”) and in person (“The security guard escorted the violator not only out of the store, but out of the mall for not wearing a mask”). Other cases began with (one of the) victim(s) acting, with punishment ultimately falling to a third party (“A neighbour stole some parcels from our block of flats. Another neighbour caught them when she was looking for her parcel. She filmed the thief and then reported them to the management company. The letting agency started eviction proceedings”).*

The participants’ levels of involvement also varied, from being actively involved as a violator *(“In response to the wrong behaviour (parking a car too close to another car), I was shouted at by a family member and received a verbal telling off”),* witnessing an incident *(“There was a gathering next door to me and a couple of the teens came out onto the street having an argument. The girl was shouting at the boy for paying a lot of attention to another girl and not her, so she said she wouldn't give him a lift home and he had to make his own way home”),* or being told about it after the fact *(“The individual told me they had to listen to inappropriate comments from a taxi driver whilst they were being driven somewhere and they felt offended by this. As a punishment, the punisher determined to leave a negative review and put in a complaint about the driver”).*

Included were implicit judgements of justness *(“The punisher terminated the employees employment. The victim had made a small error at work”),* and more explicit explanations of severity *(“It's relatively mundane, but also pretty serious too at the same time. A member of staff at our business was caught mopping the benches of a food preparation work place area with the same dirty mop that they used to clean the floors with. They were verbally berated by one of the managers and will face further action. The verbal confrontation amounted to a humiliation”).*

Punishments occurred across settings, from online spaces *(“An influencer that I follow on Social Media posted an 'ad' for Slimfast, which seemed very 'off brand' and wasn't her usual style. Some followers took umbrage to this, and called her out on the post and punished her vocally - ie asking her to explain herself, advising that she had crossed a line with her followers, asking for more research or back up to why she is using these Slimfast products and promoting them online to, what could be, a vulnerable audience”)* to in-person situations *(“A social gathering took place in a house. A young lady was reminded not to drink too much as their behaviour changes with alcohol. However they did drink to excess and became loud and lewd. The punishment was they will be excluded from the next gathering”).* Sometimes, entire groups

rather than individuals were punished (*“Some members of the art society broke the University's rules by drinking alcohol in a University building. As a result, the University inflicted a ban on the whole society from booking any University rooms in order to hold events, for the rest of the academic year”*).

Overall, these responses reflect the nuanced ways participants described punishing behaviours they encountered in everyday life.

# Appendix C

## Supplementary Information for Chapter 4

### C1 Pilot Studies on Meta-Accuracy and Bias

I conducted two pilot studies as projects with 32 (Pilot Study 1) and 45 (Pilot Study 2) second-year students. These students provided some input into the study design, and collected the data. In Pilot Study 1, vignettes depicting situations in which someone managed and did not manage their reputation were used to assess agreement between those imaging themselves as actors (providing meta-perceptions) and judges (providing actual perceptions). In Pilot Study 2 we solicited real life instances of recent selfish behaviours with accompanying meta-perceptions and subsequent actual judgements provided by observer participants. Both studies found that actors were meta-sensitive (i.e., actors displayed meta-perceptive accuracy) whilst simultaneously exhibiting a negative meta-bias (i.e., actors believed they would be judged less positively than they actually were).

#### Pilot Study 1

Humans are a highly cooperative species. One explanation for this is concern for reputation, specifically the desire to gain a positive or avoid a negative reputation (Barclay, 2012). Individuals use reputation management to influence how others view them (Dores Cruz et al., 2020), e.g. by concealing information (Piazza & Bering, 2010), or by adjusting their behaviour (Wu et al., 2016). To effectively influence how they are perceived by others, individuals must hold accurate meta-perceptions. Meta-perceptions are the beliefs we hold about how others perceive us (Laing et al., 1966), while meta-accuracy is the degree to which our meta-perceptions are correct (Kenny & DePaulo, 1993). People are aware that others do not see them as they see themselves (Carlson & Kenny, 2012; Carlson et al., 2011), and are fairly accurate at estimating how others – both strangers and acquaintances - perceive them (Carlson et al., 2010; Carlson et al., 2011; Kenny & DePaulo, 1993; Shechtman & Kenny, 1994). Individuals can differentiate how they are perceived across social contexts (Carlson & Furr, 2009), and are even quite good at predicting how others will see them when they unsuccessfully try to manage their reputation (Wu & Zheng, 2019).

Whenever not perfectly accurate, individuals can either err by (i) overestimating how favourably others view them (i.e. think others see them more favourably than they actually do), or (ii) underestimate how favourably others view them (i.e. think others see them less favourably than they actually do). As one cannot eliminate both errors at the same time, making fewer errors of one type results in more errors of the other. When facing asymmetric error costs, meaning when two errors have different cost magnitudes, people should be more likely to commit cheap errors to reduce the likelihood of engaging in expensive errors (Efferson et al., 2020; McKay & Efferson, 2010). In this case, erring on the safe side would plausibly be to underestimate how positively one is regarded (negativity bias) to remain motivated to monitor and manage one's behaviour and reputation, avoiding the consequences of a negative reputation (Barranti et al., 2016; FeldmanHall & Shenhav, 2019).

Therefore, we expected actor participants to exhibit a negative meta-bias. Simultaneously, we predicted that participants would be able to predict how they would be viewed for various (hypothetical) cooperative and uncooperative behaviours, exhibiting meta-accuracy. To investigate metacognitive accuracy and bias in predicting how others will judge us for our cooperative/non-cooperative behaviours, Pilot Study 1 used the agreement between (i) actors' perceptions of how they believe they would be judged for their (hypothetical) actions and (ii) actual judgments. Behaviours were based on common situations in which we manage and fail to manage our cooperative reputation in daily life.

## **Methods**

### **Participants**

135 participants (92 female, 43 male) aged between 18 and 62 years ( $M = 24.44$ ,  $SD = 11.36$ ) were recruited by undergraduate students from their pool of friends and family members. Participants were predominantly Asian (37%) or White (35%) and from the UK (35%), Turkey (10%), China (10%) or Egypt (10%). 61 of these participants were recruited as actor, and 74 as observer participants.

### **Materials and Procedure**

In groups, second year students were tasked with creating scenarios which they believed best tap into cooperative reputation management behaviours. These suggestions were collated before all students voted on which scenarios they considered to be the most appropriate. The



10 scenarios which received the most votes, and an additional scenario which was suggested by multiple groups, were selected as stimuli. Each scenario consisted of two items: one in which reputation is managed (cooperative behaviour) and one in which reputation is not managed (uncooperative behaviour), resulting in 22 items. Examples of these scenarios include giving versus not giving your seat to an elderly lady on the bus, collecting versus leaving your rubbish after a picnic, keeping found cash versus handing it in at a customer service desk, having a loud or quiet phone call in public, or ordering a cheaper versus a more expensive version of your usual meal when a friend is paying.

The study was built in Qualtrics, and participants were informed that their task would be to make judgments about how they think certain everyday actions would affect reputation. Subsequently, participants were randomly assigned to either imagine themselves as the actor or an observer. After providing demographic information, participants were informed that they would be ‘presented with multiple scenarios about everyday behaviours’ and that their task was to imagine themselves / another person engaging in these actions, and to (i) indicate how they think other people would judge them for it (actors) or (ii) to indicate how they would judge the actor for the behaviours (observers). Responses ranged from 1 = “Extremely Negatively” to 7 = “Extremely Positively”.

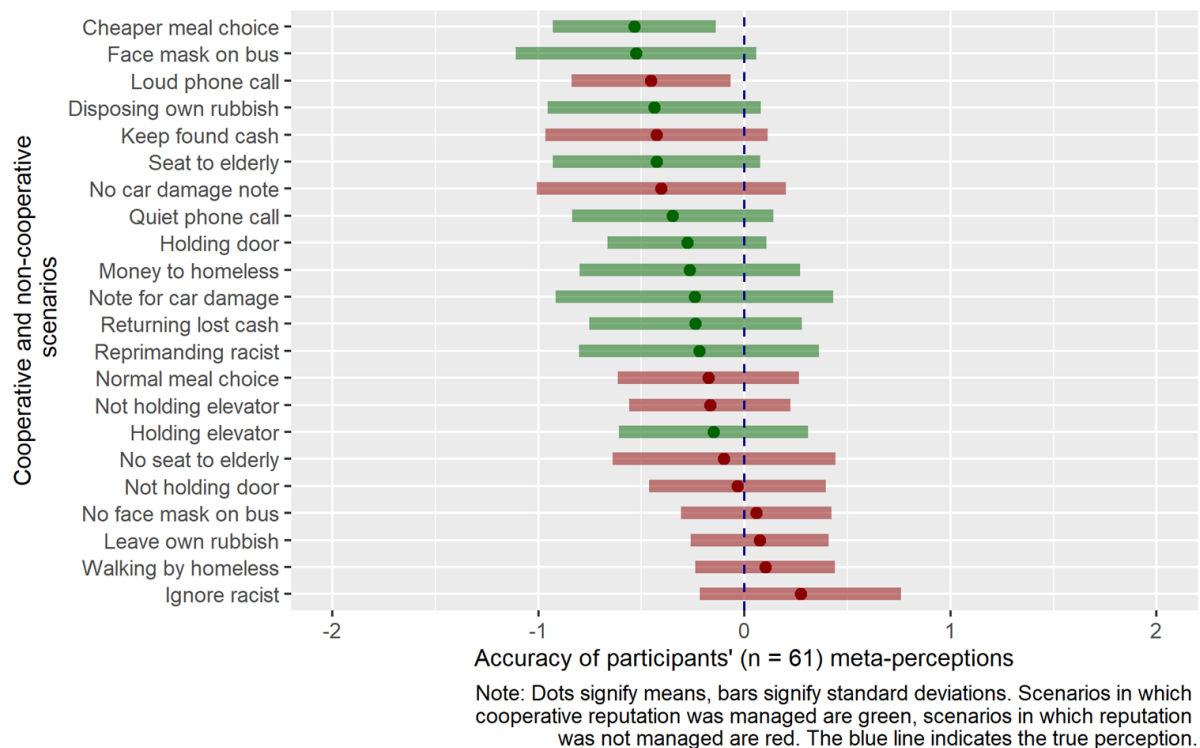
Following this, participants completed the Single-Item Narcissism Scale (Konrath et al., 2014), the 8-item Empathy Quotient (Loewen et al., 2009) on a 4-point scale from “strongly agree” to “strongly disagree”, and the 12-item Shortened Social Interaction Anxiety Scale and Social Phobia Scale (SIAS-6 & SPS-6; Peters et al., 2012) on a 5-point scale from “not at all characteristic or true of me” to “extremely characteristic or true of me” in random order. A debrief completed the survey.

## Results

To test whether actor participants exhibited meta-bias, actor accuracy scores were computed by, per item, subtracting the average of observers actual evaluations from the actors’ prediction. An intercept-only mixed-effects model with participant and item as random effects with random intercepts revealed that actors systematically expected more negative evaluations than they actually received ( $b = -0.23$ ,  $SE = 0.05$ ,  $t(19.58) = -4.78$ ,  $p = .00012$ ). See Figure C1.1 for a visualisation of this negative meta-bias.

**Figure C1.1**

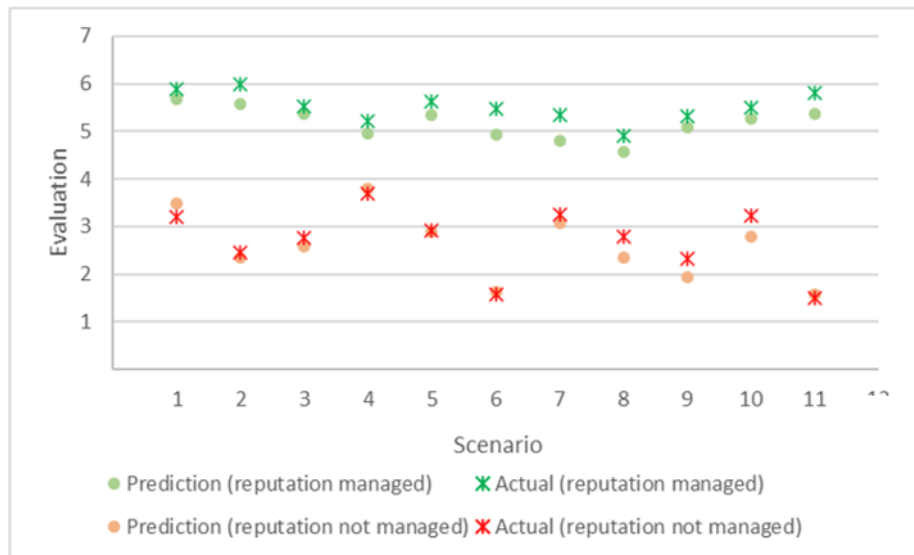
*Actor participants' meta-bias per item.*



Simultaneously, actor predictions were positively correlated with actual judgements ( $r(20) = .99, p < .001$ ). Figure C1.2 visualises the significant agreement between predictions and actual evaluations. No individual difference moderators of meta-accuracy were found, as accuracy did not significantly vary across gender, age or ethnicity, nor correlate with narcissism, empathy or social interaction anxiety ( $p > .05$ ).

**Figure C1.2**

*Agreement between average actor predictions and average actual judgements across all scenarios.*



As predicted, participants exhibited meta-sensitivity, as they could accurately anticipate how they would be viewed for certain actions relative to other actions. At the same time, they systematically underestimated how positively they would be viewed for their actions (negative meta-bias). However, these scenarios were hypothetical and actor participants did not actually engage in the behaviours. Therefore, actor participants may also have imagined how they would judge someone else for engaging in the behaviour and use that as their prediction for how others would evaluate them rather than properly imagining themselves as the actor. To improve ecological validity, subsequent projects therefore solicited real-life instances of recent behaviours, and additionally investigated more facets of an actors cooperative reputation than how positively or negatively they were perceived.

## Pilot Study 2

As an extension of Pilot Study 1, participants ( $N = 200$ ) in Pilot Study 2 were asked to either (i) self-report a recent selfish behaviour they engaged in and to predict how others would judge them for their behaviour or (ii) evaluate prior participants based on their recent selfish behaviours. Predictions and judgements were focused on the reputational consequences of actions (specifically the perceived trustworthiness, honesty, reliability, likeability and competence of actors) and motivations. Additional individual difference measures were used to

explore whether personality traits may influence metacognitive accuracy (e.g., socially anxious individuals were predicted to expect harsher judgments, and narcissistic individuals to believe they are viewed more positively than they actually are).

## **Methods**

### **Participants and Design**

Actor participants ( $n = 79$ ) had a mean self-reported age of 24 years ( $SD = 10.86$ , range: 18-60 years), with 67% identifying as female and 30% as male. The sample was predominantly comprised of individuals of British (44%) or Chinese (24%) nationality, with Asian (41%) and White (34%) ethnicities being the most common. Observer participants ( $n = 121$ ) had a mean self-reported age of 27 years ( $SD = 13.07$ , range: 18-66), with 69% identifying as female and 31% as male. Over half were of British (55%) and many others of Chinese (20%) nationality, again with Asian (45%) and White (44%) ethnicities being the most common. All participants provided informed consent before participating and the research was approved by the UCL Research Ethics Committee (Project ID: ICN-PWB-NR-23-10-20A). Observers read a random selection of 5 ‘stories’ (i.e., descriptions of a recent selfish behaviour by actor participants). 19 target stories had previously been excluded from further analysis for not describing a selfish behaviour or for being unclear, e.g. due to containing too little information. This left 60 stories to be rated by around 10 observers each. The median duration for actor participants to complete the experiment was 6 minutes 30 seconds, and 4 minutes 28 seconds for observers.

### **Procedure**

Students voted on the specific behaviour related to cooperative reputation they wanted to investigate, and additionally decided on which secondary measures they wanted to include in the study.

Data collection took place in two stages. First, actor participants were told that they would be asked to describe a recent selfish behaviour, and to predict how others will think of them based on that description. They were aware that we would ask another set of participants to read their behaviour description, and to indicate how they view them based on that behaviour. Further, they were informed that “In our daily lives we sometimes behave selfishly, for various reasons.

Examples include being unwilling to compromise, not fulfilling our share of responsibilities, or not helping someone.” before being asked to describe, in 2-4 sentences, a recent instance in which they behaved selfishly. After providing their behaviour description, actor participants were asked to imagine that someone else – the average person – reads their described behaviour and has to rate how they view them based on it. The task was to indicate how they think others would rate their traits based on their described behaviour. The behaviour was repeated at the top of the page. The items were “How trustworthy [reliable / honest / likeable / competent] would someone else rate you, based on your behaviour?” with answers ranging from 1 = “Very Untrustworthy [Unreliable / Dishonest / Unlikeable / Incompetent]” to 7 = “Very Trustworthy [Reliable / Honest / Likeable / Competent]”, with 4 = “Neutral” as midpoint. Another item asked “To what extent would they think your behaviour was driven by selfish or selfless motivations?” from 1 = “Very Selfish” to 7 = “Very Selfless”. Finally, participants completed the Single-Item Narcissism Scale (Konrath et al., 2014) and the 12-item Shortened Social Interaction Anxiety Scale and Social Phobia Scale (SIAS-6 & SPS-6; Peters et al., 2012) in counterbalanced order before providing their demographic information.

19 of the 79 collected behaviour descriptions were excluded from further analysis for being unclear or for not describing a selfish behaviour. The remaining 60 ‘stories’ were provided as stimuli for observer participants. Observers were informed that the study aims to investigate whether people can accurately predict how others will judge them based on their real-life behaviours, with them taking part in the second stage. Specifically, their task was to read descriptions provided by five prior participants, and to indicate how they perceive those participants based on their behaviour. Observers were given the same information on selfish behaviours as actor participants were before being shown a recent selfish behaviour a prior participant engaged in. Finally, along the same items as provided to actors, observers indicated how they perceive the author based on their behaviour description.

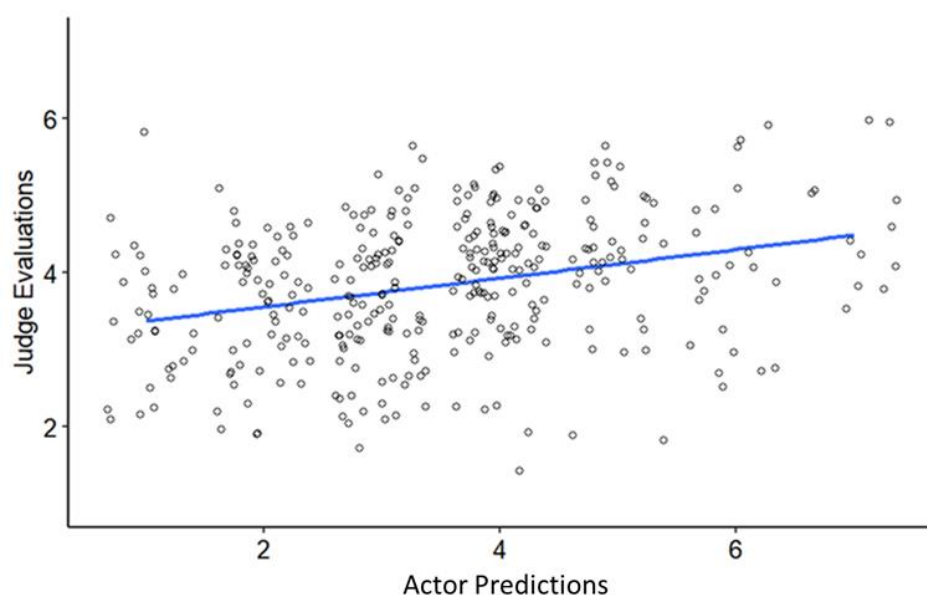
## Results

With the agreement between actor predictions and judge evaluations determining meta-accuracy, we found that participants were meta-sensitive and could generally successfully estimate how they would be judged for their recent selfish actions,  $r(345) = .30, p < .0001, 95\% \text{ CI } [0.20, 0.40]$  (Figure C1.3). However, when reputational items were considered in isolation,

actors were only statistically significantly meta-accurate when predicting perceptions of honesty,  $r(58) = .39, p = .002, 95\% \text{ CI } [0.15, 0.57]$ .

**Figure C1.3**

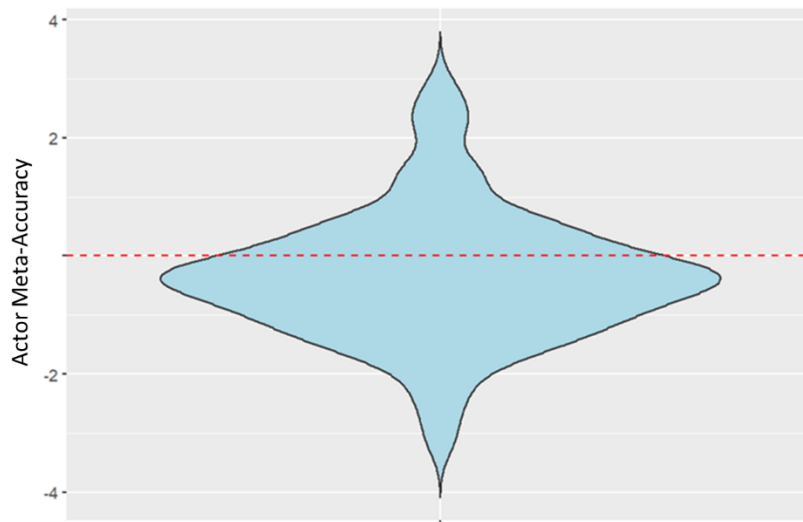
*Meta-Accuracy: The agreement between actor predictions and judge evaluations*



Averaged actor accuracy scores ranged from -3.09 to 2.80 with a mean of -0.35 ( $SD = 1.10$ ). In line with this, when predictions and actual evaluations differed, participants exhibited metacognitive bias and tended to overestimate the negative effects of acting selfishly on their reputation,  $t(59) = -2.45, p = .017, 95\% \text{ CI } [-0.63, -0.06]$  (Figure C1.4). This may be understood in terms of error-management, where costlier error types tend to be minimised (Efferson et al., 2020; McKay & Efferson, 2010). When checking whether meta-bias is especially present for some reputational aspects, a statistically significant negative meta-bias was present for honesty ( $t(59) = -3.38, p = .001, 95\% \text{ CI } [-0.99, -0.25]$ ), trustworthiness ( $t(59) = -2.49, p = .016, 95\% \text{ CI } [-0.76, -0.08]$ ) and likeability ( $t(59) = -2.25, p = .028, 95\% \text{ CI } [-0.84, -0.05]$ ).

**Figure C1.4**

*Systematic Negative Meta-Bias: Violin plot with red line through zero indicating where perfect meta-accuracy (no bias) lies*



## Discussion

In both studies, participants exhibited both meta-sensitivity and a negative meta-bias. I expected to find the same in the main study investigating the meta-perceptive accuracy of cooperative reputations. The main study is more complex than the pilot studies. First, the pilot studies solicited only meta- and observer-perceptions whereas the main study also solicited self-perceptions. Secondly, the pilot studies include fewer items with only the second pilot study including a single motivation item at all. Due to these changes, the main study could assess not only actor meta-accuracy, but also observer accuracy. Moreover, more advanced analysis methods such as a componential approach and more sophisticated models (e.g. baseline, distinctive, and normative accuracy, with insight and transparency bias) could be applied.

## **C2 Preregistration for “Meta-Cognitive Accuracy in Assessing the Influence of Cooperative and Uncooperative Actions on Reputation”**

### **Meta-Accuracy around Cooperation (#159632)**

#### **Author(s)**

**Pre-registered on:** 01/26/2024 01:49 AM (PT)

This pre-registration is currently anonymous to enable blind peer-review.

It has one author.

#### **1) Have any data been collected for this study already?**

No, no data have been collected for this study yet.

#### **2) What's the main question being asked or hypothesis being tested in this study?**

In the complexity of human social interactions, the interplay between self-perception and the perceptions of others holds a pivotal role. Central to this dynamic are meta-perceptions - our beliefs about how others perceive us - and the degree to which these meta-perceptions align with reality, known as meta-accuracy. We will investigate whether people know how their cooperative and uncooperative actions are viewed by others, and whether other people make accurate judgements of an actor's motives upon learning about specific behaviours. Our main predictions are, that:

- Participants will provide accurate meta-perceptions and observer judgments for both cooperative and uncooperative behaviours.
- Meta-accuracy and observer accuracy will be lower for uncooperative actions than for cooperative actions.
- Actors will exhibit a negative meta-bias.

#### **3) Describe the key dependent variable(s) specifying how they will be measured.**

Actor participants will describe a recent uncooperative and a recent cooperative behaviour they engaged in, each with accompanying self-perceptions and meta-perceptions (i.e., predictions of how they believe others will view them based on their behaviour). Observer participants will then provide actual judgements on a random sub-set of those behaviour descriptions. Self-perceptions, meta-perceptions, and actual judgements will be given for the same 15 motive and 10 trait items, all rated from 1 = "Not at all" to 7 = "Completely" (i.e., 25 ratings each for meta-perceptions, self-perceptions and actual judgments per behaviour). Estimates provided by participants (i.e. meta-perceptions for actor accuracy and judgements for observer accuracy) will be modelled as the response variables, and the true values (i.e. observer judgements for actor accuracy and actor self-perceptions for observer accuracy) will be modelled as predictors. We will not run analyses directly with Likert-scale responses, and will instead use cluster centring within actor to ensure that perceptions are anchored to the mean validation criterion within each actor (for more detail see end of Question 5).

#### **4) How many and which conditions will participants be assigned to?**

Data collection will occur in two stages: first we will recruit actor participants who will each provide one description of a recent cooperative AND one description of a recent uncooperative behaviour they engaged in, with accompanying self- and meta-perceptions. In a second stage,



observers will read a random subset of these behaviour descriptions (4 uncooperative and 4 cooperative behaviours) and provide judgements.

**5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.**

We will fit mixed-effects models when investigating meta-accuracy and observer accuracy.

Models will be fit separately for cooperative and uncooperative behaviours.

Baseline accuracy models test for a linear relationship between estimated and true values in actor meta-perceptions (meta-perception ~ observer judgment) and observer attributions (observer judgment ~ self-perception). However, they do not control for potential confounds of judgment accuracy: meta-accuracy depends both on an understanding of how one is uniquely perceived (distinctive meta-accuracy) and how the average actor is perceived (normative accuracy). The second models will therefore test for distinctive accuracy by modelling the relationship between the estimated value and the normative profile on top of the baseline models' relationship between the estimated and true values. The normative profile is the grand mean centred distribution of the self-perception means across actors. The third models will test for insight. Actor insight is the relationship between meta-perceptions and judgements while controlling for self-perceptions. As the model controls for self-perception, the relationship between meta-perceptions and actual judgements can be interpreted as actors' understanding of how they will be misperceived. The relationship between meta-perception and self-perception shows transparency bias – the extent to which actors overestimate how accurately they will be perceived. Observer insight is the relationship between judgements and self-perceptions while controlling for meta-perceptions.

To examine systematic biases in meta-perceptions we will variable-centre the data and interact item with observer judgment in the baseline meta-accuracy model (meta-perceptions centred on the mean observer judgment of each item & observers' judgments on the mean self-perception of each item). Values of zero represent average point estimate accuracy by item (i.e., estimates which differ from zero show systematic directional bias).

We will adopted a maximal random structures approach, with random slopes and intercepts for all linear predictors across all random effects. Whenever a singular fit is obtained we will remove the random slope with the smallest standard deviation. When predicting actor meta-accuracy a single random effect (actor) will be specified, and when predicting observer accuracy two crossed random effects will be included (actor & observer). We will use cluster centring within actor to ensure that meta- and observer-perceptions are anchored to the mean validation criterion within each actor (i.e., meta-perceptions will be true-mean centred on the average observer judgment towards an actor, observer attributions will be true-mean centred on the average self-perception provided by actors, and actors' self-perceptions will be mean-centred within actor).

**6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.**

We will exclude any participant who completes the study 'exceptionally fast' (3 standard deviations below the mean). Actor participants who did not describe a cooperative/uncooperative behaviour or were unclear (e.g., due to containing too little information) will be removed. We will also exclude judges who incorrectly answer at least one out of two attention check questions.

**7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.**

The sample size is derived it from Lees et al. (2022). Their fourth study – the study which is most similar to the present study - produced a significant moderator of meta-accuracy (our most complex relevant model). Using the "mixedpower" package in R we simulated power for the significant moderator of meta-accuracy model. This data-based estimation used the beta coefficients found in the Study 4 data. Further, actor participant was selected as the random effect to vary in the simulation, with steps between 60-150 as the different sample sizes power should be simulated for. The critical t-value was set to 2, which reflects an alpha level of 5%. The results suggest, that 140 actor participants are sufficient to detect a moderating effect of meta-accuracy of similar size (simulated power = 80%, based on 1000 simulations). We will therefore aim to collect 140 usable cooperative behaviours descriptions and 140 usable uncooperative behaviour descriptions. We will oversample by 40 participants, anticipating the need for exclusions, resulting in an aim to recruit 180 actor participants.

The sample size for observers is based on the decision to have at least 10 judges rate each behaviour description, and an additional trade-off between a high number of observations and time (i.e., to reduce attrition and attention loss we aim to limit the survey time to 15 minutes). Each observer will therefore be asked to rate 8 stories, meaning that 350 observers are needed so that each of the anticipated 280 descriptions (140 for each behaviour) will be rated by around 10 observers. In anticipation of exclusions, we plan to recruit 380 observer participants.

**8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?) Additional Predictions:**

- We will investigate several potential trait moderators of (a) meta-accuracy and (b) meta-bias.
- Participants will demonstrate higher meta-accuracy and observer judgement accuracy for relevant, observable items (e.g. cooperativeness) compared to internal (e.g. motivation items) or less relevant items (e.g. competence).
- Both actors and judges will perceive accounts of uncooperative behaviour as more indicative of actor character than accounts of cooperative behaviour, but they will expect predictions and actual judgements to align more for cooperative compared to uncooperative behaviours.
- Some items will be seen as distinctly applicable to uncooperative versus cooperative behaviours.

**Analysis:**

- Accuracy moderators: Building on the distinctive accuracy model, we will interact the moderator (i.e., mean-centred status, mean-centred concern for reputation, mean-centred social perspective taking, item, or condition) with distinctive accuracy and normative accuracy.

- In three linear models, we will test for the moderating effect of neuroticism, narcissism and concern for reputation on meta bias by examining the relationship between actors' meta-perceptive accuracy scores (calculated as the difference between actor meta-perception and actual observer judgments, both of which will be mean-centred within item) and their mean-centred neuroticism, narcissism, or concern for reputation scores. To additionally test for the potential effect of behaviour type condition (uncooperative or cooperative), moderators will be interacted with condition.

- To assess whether participants exhibit higher distinctive accuracy on certain items, we will perform pairwise comparisons on the estimated marginal means for the interaction between item and distinctive accuracy, while adjusting for multiple comparisons using the Tukey method.

For the models assessing the moderating role of uncooperative versus cooperative behaviour on meta-perceptive or judgment accuracy the data sets for the two behaviours will be combined.

-To examine (i) whether participants anticipate greater alignment between predictions and actual judgements for uncooperative or cooperative behaviours and (ii) whether they consider uncooperative or cooperative behaviours to be more indicative of actor character, ratings for those questions will be centred around the scale midpoint of 4. These adjusted ratings will then be employed as the response variables in four constant (intercept-only) linear models.

-To assess whether certain items are seen as distinctly applicable to uncooperative versus cooperative behaviours, we will conduct mixed-effects ANOVAs and perform pairwise comparisons on the estimated marginal means for the interaction between item and condition, while adjusting for multiple comparisons using the Tukey method.

**Available at** <https://aspredicted.org/xdg9-68hz.pdf>

Version of AsPredicted Questions: 2.00

## C3 Study Materials (Qualtrics Surveys for Actors and Observers)

### Meta-Cooperative Accuracy Stage 1 (Actor)

---

#### Start of Block: info & consent

info *This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, you will be asked to describe a recent cooperative and uncooperative behaviour you have engaged in, before predicting how others might judge you based on those behaviours.**

The study is estimated to last around 23 minutes.

Your submission will be rejected if you do not follow the task instructions when providing behaviour descriptions or if you complete the study exceptionally fast. There is also an attention check.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. We will collect no identifying information. Completely de-identified datafiles may be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return the survey.

---

#### Page Break

#### notice **Local Data Protection Privacy Notice**

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: 'Public task'

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return the survey.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)

---

consent Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can message the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
- (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
- (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).
- (5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning the study on Prolific.

☐ I **consent** to take part in this study (1)

☐ I do **NOT consent** to take part in this study (2)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent</strong> to take part in this study*

Page Break



ID Please enter your Prolific ID

Page Break

consent 2

**PLEASE READ THE TASK DESCRIPTION CAREFULLY**

This study aims to investigate whether people can accurately predict how others will judge them based on their real-life behaviours. For this, we will conduct a study in two stages. **You are invited to take part in the first stage** of this study.

As part of this study we will ask you to **describe a recent cooperative and a recent uncooperative behaviour**, before **predicting how others might think of you** based on each description. In the second stage of the study, we will ask another set of participants to read your behaviour descriptions, before indicating how they view you based on the behaviour.

Therefore, we would like to use the descriptions you provide as stimuli for participants to rate in future studies. Your behaviour descriptions would only be used for scientific purposes and would be completely anonymised.

***Do you allow us to use your descriptions in future research?***

*Note: If you do not consent, please return the study*

☐ I **consent** (1)

☐ I do **NOT** consent (2)

*Skip To: End of Survey If PLEASE READ THE TASK DESCRIPTION CAREFULLY This study aims to investigate whether people can accu... = I do <strong>NOT </strong>consent*

End of Block: info & consent

---

Start of Block: coBeh

coBehaviour In our daily lives we sometimes behave cooperatively or uncooperatively, for various reasons. Examples of uncooperative behaviours include selfish actions such as not pulling your weight, littering, or being unhelpful. Examples of **cooperative** behaviours include prosocial actions such as helping or collaborating well with others, donating blood, and sharing.

**Please describe in 2-5 sentences, ONE recent instance in which you behaved *cooperatively*.**

After providing your description, you will be asked to provide **self-perceptions** AND **predictions for how others will evaluate you** for this cooperative behaviour. Please do not include any identifying information in your writing. However, *please be clear and specific enough so that future participants will understand your description. Providing context is encouraged.*

Note that future participants will see the behaviour exactly as you describe it below, without any changes made.

---

---

End of Block: coBeh

---

Start of Block: CO meta

coMetaTrait Imagine *someone else* – the average person – reads your described behaviour (as shown below) and has to rate how they view you based on this. Please indicate how you think **they** would rate your traits based on this described behaviour.

The description: “*coBehaviour/ChoiceTextEntryValue*”

|                     | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|---------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| trustworthy<br>(1)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| sincere (2)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| honest (3)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| loyal (4)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| just (5)            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| reliable (6)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| empathic<br>(7)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| likeable (8)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| competent<br>(9)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| intelligent<br>(10) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |

Page Break



coMetaMotive Imagine *someone else* – the average person – reads your described behaviour (as shown below) and has to rate how much they think your behaviour was motivated by various factors. Please indicate how you think **they** would rate your motivations based on your described behaviour.

The description: “*coBehaviour/ChoiceTextEntryValue*”

|                                                      | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| gaining a good /<br>avoiding a bad<br>reputation (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| selflessness (2)                                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| obligation (3)                                       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| avoiding<br>conflict (4)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| helping others<br>(5)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| fairness (6)                                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| empathy (7)                                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| averting guilt<br>(8)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| self-interest (9)                                    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| laziness (10)                                        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| pride (11)                                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| spite (12)                                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| jealousy (13)                                        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| competitiveness<br>(14)                              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| revenge (15)                                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |



End of Block: CO meta

Start of Block: CO self



coSelfTrait Based on your described behaviour (as shown below), how would **you** rate yourself on the following traits?

The description: “*coBehaviour/ChoiceTextEntryValue*”

|                  | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| trustworthy (1)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| sincere (2)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| honest (3)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| loyal (4)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| just (5)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| reliable (6)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| empathic (7)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (8)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| competent (9)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| intelligent (10) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break

coSelfMotive When engaging in your described behaviour (as shown below), how much were **you** motivated by the following factors?

The description: “*coBehaviour/ChoiceTextEntryValue*”

|                                                      | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| gaining a good /<br>avoiding a bad<br>reputation (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| selflessness (4)                                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| obligation (5)                                       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| avoiding<br>conflict (6)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| helping others<br>(7)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| fairness (8)                                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| empathy (9)                                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| averting guilt<br>(10)                               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| self-interest<br>(11)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| laziness (12)                                        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| pride (13)                                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| spite (14)                                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| jealousy (15)                                        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| competitiveness<br>(16)                              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| revenge (17)                                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |

End of Block: CO self

## Start of Block: unBeh

uncBeh In our daily lives we sometimes behave cooperatively or uncooperatively, for various reasons. Examples of **uncooperative** behaviours include selfish actions such as not pulling your weight, littering, or being unhelpful. Examples of cooperative behaviours include prosocial actions such as helping or collaborating well with others, donating blood, and sharing.

**Please describe in 2-5 sentences, ONE recent instance in which you behaved *uncooperatively*.**

After providing your description, you will be asked to provide **self-perceptions** AND **predictions for how others will evaluate you** for this uncooperative behaviour. Please do not include any identifying information in your writing. However, *please be clear and specific enough so that future participants will understand your description.* **Providing context is encouraged.**

Note that future participants will see the behaviour exactly as you describe it below, without any changes made.

---

---

## End of Block: unBeh

## Start of Block: UN meta



unMetaTrait Imagine *someone else* – the average person – reads your described behaviour (as shown below) and has to rate how they view you based on this. Please indicate how you think **they** would rate your traits based on this described behaviour.

The description: “*\${uncBeh/ChoiceTextEntryValue}*”

|                 | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|-----------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| trustworthy (4) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| sincere (7)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| honest (8)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| loyal (9)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| just (10)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                     |                       |                       |                       |                       |                       |                       |                       |
|---------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| reliable<br>(11)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| empathic<br>(12)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable<br>(13)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| competent<br>(14)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| intelligent<br>(15) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break



unMetaMotive Imagine *someone else* – the average person – reads your described behaviour (as shown below) and has to rate how much they think your behaviour was motivated by various factors. Please indicate how you think **they** would rate your motivations based on your described behaviour.

The description: “*uncBeh/ChoiceTextEntryValue*”

|                                                      | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| gaining a good /<br>avoiding a bad<br>reputation (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| selflessness (4)                                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| obligation (5)                                       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| avoiding<br>conflict (6)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| helping others<br>(7)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| fairness (8)                                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |

|                      |                       |                       |                       |                       |                       |                       |                       |
|----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| empathy (9)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| averting guilt (10)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| self-interest (11)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| laziness (12)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| pride (13)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| spite (14)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| jealousy (15)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| competitiveness (16) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| revenge (17)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: UN meta

Start of Block: UN self



unSelfTrait Based on your described behaviour (as shown below), how would **you** rate yourself on the following traits?

The description: “*uncBeh/ChoiceTextEntryValue*”

|                 | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|-----------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| trustworthy (4) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| sincere (7)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| honest (8)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                  |                       |                       |                       |                       |                       |                       |                       |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| loyal (9)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| just (10)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| reliable (11)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| empathic (12)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (13)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| competent (14)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| intelligent (15) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break



unSelfMotive When engaging in your described behaviour (as shown below), how much were **you** motivated by the following factors?

The description: “*uncBeh/ChoiceTextEntryValue*”

|                                                | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| gaining a good / avoiding a bad reputation (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selflessness (4)                               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| obligation (5)                                 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| avoiding conflict (6)                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| helping others (7)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                      |                       |                       |                       |                       |                       |                       |                       |
|----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| fairness (8)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| empathy (9)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| averting guilt (10)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| self-interest (11)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| laziness (12)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| pride (13)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| spite (14)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| jealousy (15)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| competitiveness (16) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| revenge (17)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: UN self

Start of Block: predict

accuracyPred Do you think you were more accurate in predicting how others will judge you based on your uncooperative behaviour or more accurate in predicting how others will judge you based on your cooperative behaviour?

|     | 1. Definitely Uncooperative (1) | 2 (2)                 | 3 (3)                 | 4. Equal (4)          | 5 (5)                 | 6 (6)                 | 7. Definitely Cooperative (7) |
|-----|---------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------------|
| (1) | <input type="radio"/>           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |

Page Break

characPredSelf Do you think the uncooperative or the cooperative behaviour you described was more indicative of your character?

|     | 1.<br>Uncooperative<br>Behaviour (1) | 2 (2)                 | 3 (3)                 | 4. Equal<br>(4)       | 5 (5)                 | 6 (6)                 | 7.<br>Cooperative<br>Behaviour<br>(7) |
|-----|--------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|---------------------------------------|
| (1) | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                 |

Page Break



attention Do please select uncooperative behaviour, the very first button, to show that you are paying attention.

|     | 1.<br>Uncooperative<br>Behaviour (1) | 2 (0)                 | 3 (0)                 | 4. Equal<br>(0)       | 5 (0)                 | 6 (0)                 | 7.<br>Cooperative<br>Behaviour<br>(0) |
|-----|--------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|---------------------------------------|
| (1) | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                 |

End of Block: predict

Start of Block: status



status **Think of this ladder as representing where people stand in your country.**

At the top of the ladder are people who are the **best off** – those who have the most money, the most education, and the most respected jobs. At the bottom are the people who are the **worst off** – who have the least money, the least education, and the least respected jobs or no job. The higher up you are on this ladder, the closer you are to the people at the very top; the lower you are, the closer you are to the people at the very bottom.

**Where do you think you fall on the social ladder, relative to other people?**

*Please enter a numeral between 1 and 10.*

End of Block: status



---

**Start of Block: repConcern**

repConcern Please indicate how characteristic the following statements are of you. Responses range from "Not at all characteristic of me" to "Extremely characteristic of me".

|                                                                                               | 1. Not at all<br>characteristic<br>for me (1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5. Extremely<br>characteristic<br>for me (5) |
|-----------------------------------------------------------------------------------------------|-----------------------------------------------|-----------------------|-----------------------|-----------------------|----------------------------------------------|
| I am rarely<br>concerned<br>about my<br>reputation (1)                                        | <input type="radio"/>                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                        |
| I do not<br>consider what<br>others say<br>about me (4)                                       | <input type="radio"/>                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                        |
| I wish to have a<br>good reputation<br>(5)                                                    | <input type="radio"/>                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                        |
| If my reputation<br>is not good, I<br>feel very bad (6)                                       | <input type="radio"/>                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                        |
| I find it<br>important that<br>others consider<br>my reputation<br>as a serious<br>matter (7) | <input type="radio"/>                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                        |
| I try hard to<br>work on my<br>reputation (in<br>my<br>relationships<br>with others) (8)      | <input type="radio"/>                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                        |
| I find it difficult<br>if others paint<br>an incorrect<br>image of me (9)                     | <input type="radio"/>                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                        |

---

**End of Block: repConcern**

---

**Start of Block: Social Perspective Taking**



SocPersTaking Please indicate how often you do the following. Responses range from "Almost never" to "Almost all the time".

|                                                                                                                                     | 1. Almost<br>Never (1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5. Almost all<br>the time (5) |
|-------------------------------------------------------------------------------------------------------------------------------------|------------------------|-----------------------|-----------------------|-----------------------|-------------------------------|
| How often do you try to figure out how the people around you view different situations? (1)                                         | <input type="radio"/>  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| If you are having a disagreement with your friends, how often do you try to imagine how they are feeling? (4)                       | <input type="radio"/>  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| How often do you try to look at everybody's side of a disagreement before you make a decision? (5)                                  | <input type="radio"/>  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| When you are upset at someone, how often do you try to "put yourself in his or her shoes"? (6)                                      | <input type="radio"/>  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| How often do you try to understand your friends/co-workers/acquaintances better by trying to figure out what they are thinking? (7) | <input type="radio"/>  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| Before criticizing others, how often do you imagine how you would feel if you were in their place? (8)                              | <input type="radio"/>  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| To understand your friends better, how often do you imagine how things look from their perspective? (9)                             | <input type="radio"/>  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |

End of Block: Social Perspective Taking

Start of Block: neuroticism



N Please answer each question by selecting 'Yes' or 'No' following the question. There are no right or wrong answers.

|                                                              | Yes (1)               | No (0)                |
|--------------------------------------------------------------|-----------------------|-----------------------|
| Does your mood often go up and down? (1)                     | <input type="radio"/> | <input type="radio"/> |
| Do you ever feel 'just miserable' for no reason? (16)        | <input type="radio"/> | <input type="radio"/> |
| Are you an irritable person? (17)                            | <input type="radio"/> | <input type="radio"/> |
| Are your feelings easily hurt? (18)                          | <input type="radio"/> | <input type="radio"/> |
| Do you often feel 'fed-up'? (19)                             | <input type="radio"/> | <input type="radio"/> |
| Would you call yourself a nervous person? (20)               | <input type="radio"/> | <input type="radio"/> |
| Are you a worrier? (21)                                      | <input type="radio"/> | <input type="radio"/> |
| Would you call yourself tense or 'highly strung'? (22)       | <input type="radio"/> | <input type="radio"/> |
| Do you worry too long after an embarrassing experience? (23) | <input type="radio"/> | <input type="radio"/> |
| Do you suffer from 'nerves'? (24)                            | <input type="radio"/> | <input type="radio"/> |
| Do you often feel lonely? (25)                               | <input type="radio"/> | <input type="radio"/> |
| Are you often troubled about feelings of guilt? (26)         | <input type="radio"/> | <input type="radio"/> |

End of Block: neuroticism

Start of Block: narcissism

narcissism To what extent do you agree with this statement: "*I am a narcissist.*"  
(Note: The word "narcissist" means egotistical, self-focused, and vain.)

|                             | 1. Not very<br>true of me<br>(1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Very<br>true of me<br>(7) |
|-----------------------------|----------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------------|
| I am a<br>narcissist<br>(1) | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>        |

End of Block: narcissism

Start of Block: demographics

gender What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



age What is your age in years?

---



nation What is your nationality?

---

End of Block: demographics

## Meta-Cooperative Accuracy Stage 2 (Observer)

---

### Start of Block: info & consent

info *This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, you will be asked to read descriptions about recent cooperative and uncooperative behaviours prior participants have engaged in. Your task is to indicate how you perceive the authors based on their described behaviour.**

The study is estimated to last around 12 minutes.

Your submission will be rejected if you fail the attention checks or complete the study exceptionally fast.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. We will collect no identifying information. Completely de-identified datafiles may be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return the survey.

---

### Page Break

#### notice **Local Data Protection Privacy Notice**

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: 'Public task'

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return the survey.**



consent Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can message the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
- (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
- (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).
- (5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning the survey.

- ☐ I **consent** to take part in this study (1)
- ☐ I do **NOT consent** to take part in this study (0)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent</strong> to take part in this study*

---

Page Break

---



ID Please enter your Prolific ID

---

Page Break

---

task

**PLEASE READ THE TASK DESCRIPTION CAREFULLY**

This study aims to investigate whether people can accurately predict how others will judge them based on their real-life behaviours. For this, we are conducting a study in two stages. **You are invited**

**to take part in the second stage** of this study.

In the first stage we asked participants to describe a recent uncooperative and a recent cooperative behaviour they engaged in, and to then predict how others will view them based on each description. Now, in the second stage of the study, we will ask you to **read eight descriptions provided by prior participants**, and to **indicate how you perceive those participants based on their behaviour**.

---

Page Break

task2 In our daily lives we sometimes behave cooperatively or uncooperatively, for various reasons. Examples of uncooperative behaviours include selfish actions such as not pulling your weight, littering, or being unhelpful. Examples of cooperative behaviours include prosocial actions such as helping or collaborating well with others, donating blood, and sharing.

Previous participants were asked to describe a recent instance in which they behaved uncooperatively and a recent instance in which they behaved cooperatively.

End of Block: info & consent

---

Start of Block: co subset



coStory cooperative behaviours

*[actor responses]*

End of Block: co subset

---

Start of Block: CO rate



co trait Below is a description of a cooperative behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, **indicate how you perceive the author**.

“\${Im://Field/1}”

|                  | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| trustworthy (4)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| sincere (7)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| honest (8)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| loyal (9)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| just (10)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| reliable (11)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| empathic (12)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (13)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| competent (14)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| intelligent (15) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break



co motive Below is a description of a cooperative behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, **indicate what you believe the author's motives were.**



“\${Im: //Field/1}”

|                                                      | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| gaining a good /<br>avoiding a bad<br>reputation (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| selflessness (4)                                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| obligation (5)                                       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| avoiding<br>conflict (6)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| helping others<br>(7)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| fairness (8)                                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| empathy (9)                                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| averting guilt<br>(10)                               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| self-interest<br>(11)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| laziness (12)                                        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| pride (13)                                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| spite (14)                                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| jealousy (15)                                        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| competitiveness<br>(16)                              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| revenge (17)                                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |

End of Block: CO rate

Start of Block: un subset



unStory uncooperative behaviours

[actor responses]

End of Block: un subset

Start of Block: UN rate



un trait Below is a description of an uncooperative behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, **indicate how you perceive the author.**

“ $\{lm://Field/1\}$ ”

|                  | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| trustworthy (4)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| sincere (7)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| honest (8)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| loyal (9)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| just (10)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| reliable (11)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| empathic (12)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (13)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| competent (14)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| intelligent (15) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |



Below is a description of an uncooperative behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, **indicate what you believe the author's motives were.**

“*{Im://Field/1}*”

|                                                      | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| gaining a good /<br>avoiding a bad<br>reputation (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| selflessness (4)                                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| obligation (5)                                       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| avoiding<br>conflict (6)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| helping others<br>(7)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| fairness (8)                                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| empathy (9)                                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| averting guilt<br>(10)                               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| self-interest<br>(11)                                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| laziness (12)                                        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| pride (13)                                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| spite (14)                                           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| jealousy (15)                                        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |

|                      |                       |                       |                       |                       |                       |                       |                       |
|----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| competitiveness (16) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| revenge (17)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: UN rate

Start of Block: predict

better predict Do you think your judgments align more closely with previous participants' expected evaluations of their *uncooperative* behaviours, or do you think your judgments align more closely with participants' expected evaluations of their *cooperative* behaviours?

|     |                                 |                       |                       |                       |                       |                       |                               |
|-----|---------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------------|
|     | 1. Definitely Uncooperative (1) | 2 (2)                 | 3 (3)                 | 4. Equal (4)          | 5 (5)                 | 6 (6)                 | 7. Definitely Cooperative (7) |
| (1) | <input type="radio"/>           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |

Page Break

you charac Do you think the uncooperative or the cooperative behaviours described were more indicative of people's character?

|     |                                 |                       |                       |                       |                       |                       |                               |
|-----|---------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------------|
|     | 1. Uncooperative Behaviours (1) | 2 (2)                 | 3 (3)                 | 4. Equal (4)          | 5 (5)                 | 6 (6)                 | 7. Cooperative Behaviours (7) |
| (1) | <input type="radio"/>           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |

Page Break



attention Do please select uncooperative behaviours, the very first button, to show that you are paying attention.

|     | 1.<br>Uncooperative<br>Behaviours (1) | 2 (0)                 | 3 (0)                 | 4. Equal<br>(0)       | 5 (0)                 | 6 (0)                 | 7.<br>Cooperative<br>Behaviours<br>(0) |
|-----|---------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------------------------------|
| (1) | <input type="radio"/>                 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                  |

End of Block: predict

Start of Block: attention



check On which topic were you asked to provide ratings?

- ☐ cooperative / uncooperative behaviours (1)
- ☐ holiday destinations / staycations (0)
- ☐ dietary preferences / allergies (0)
- ☐ consumer behaviour / customer manipulation (0)

End of Block: attention

Start of Block: demographics

gender What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



age What is your age in years?

---



nation What is your nationality?

---

End of Block: demographics

---

## C4 Correlation Matrices of Traits and Motives for Cooperative and Uncooperative Actions

**Table C4.1**

*Correlation Matrix of Traits and Motives for Cooperative Behaviours*

|                    | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    | 11    | 12    | 13   | 14    | 15    | 16    | 17    | 18   | 19   | 20   | 21   | 22   | 23   | 24   | 25 |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|------|------|------|------|------|------|------|----|
| 1. trustworthy     | --    |       |       |       |       |       |       |       |       |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 2. sincere         | 0.62  | --    |       |       |       |       |       |       |       |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 3. honest          | 0.66  | 0.59  | --    |       |       |       |       |       |       |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 4. loyal           | 0.59  | 0.49  | 0.47  | --    |       |       |       |       |       |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 5. just            | 0.45  | 0.31  | 0.47  | 0.41  | --    |       |       |       |       |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 6. reliable        | 0.65  | 0.53  | 0.58  | 0.56  | 0.44  | --    |       |       |       |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 7. empathic        | 0.2   | 0.34  | 0.2   | 0.24  | 0.22  | 0.26  | --    |       |       |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 8. likeable        | 0.5   | 0.55  | 0.49  | 0.44  | 0.35  | 0.47  | 0.34  | --    |       |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 9. competent       | 0.52  | 0.37  | 0.53  | 0.5   | 0.44  | 0.56  | 0.03  | 0.3   | --    |       |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 10. intelligent    | 0.53  | 0.43  | 0.52  | 0.48  | 0.36  | 0.47  | 0.07  | 0.41  | 0.56  | --    |       |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 11. reputation     | 0.15  | 0.11  | 0.15  | 0.2   | -0.04 | 0.18  | -0.04 | 0.12  | 0.12  | 0.15  | --    |       |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 12. selflessness   | 0.23  | 0.23  | 0.18  | 0.21  | 0.18  | 0.23  | 0.02  | 0.21  | 0.17  | 0.25  | 0.05  | --    |      |       |       |       |       |      |      |      |      |      |      |      |    |
| 13. obligation     | 0.19  | 0.13  | 0.07  | 0.19  | 0.12  | 0.25  | -0.01 | -0.01 | 0.26  | 0.19  | 0.33  | 0.24  | --   |       |       |       |       |      |      |      |      |      |      |      |    |
| 14. avoid conflict | 0.12  | 0.0   | 0.19  | 0.09  | 0.15  | 0.19  | -0.07 | -0.05 | 0.22  | 0.12  | 0.25  | -0.02 | 0.39 | --    |       |       |       |      |      |      |      |      |      |      |    |
| 15. helping        | 0.21  | 0.33  | 0.24  | 0.19  | 0.08  | 0.21  | 0.26  | 0.36  | 0.1   | 0.09  | -0.01 | 0.2   | 0.04 | -0.18 | --    |       |       |      |      |      |      |      |      |      |    |
| 16. fairness       | 0.28  | 0.26  | 0.34  | 0.36  | 0.34  | 0.26  | 0.12  | 0.29  | 0.25  | 0.25  | 0.12  | 0.18  | 0.08 | 0.17  | 0.12  | --    |       |      |      |      |      |      |      |      |    |
| 17. empathy        | 0.22  | 0.37  | 0.29  | 0.23  | 0.21  | 0.22  | 0.64  | 0.4   | 0.11  | 0.09  | 0.2   | 0.07  | 0.08 | -0.07 | 0.38  | 0.23  | --    |      |      |      |      |      |      |      |    |
| 18. guilt          | -0.18 | -0.21 | -0.15 | -0.06 | -0.05 | -0.1  | -0.17 | -0.25 | 0.04  | -0.09 | 0.23  | 0.09  | 0.23 | 0.35  | -0.09 | 0     | -0.05 | --   |      |      |      |      |      |      |    |
| 19. selfishness    | 0.1   | 0.06  | 0.09  | 0.14  | 0.02  | 0.08  | -0.1  | -0.06 | 0.17  | 0.28  | 0.3   | 0.05  | 0.18 | 0.24  | -0.06 | 0.17  | -0.03 | 0.04 | --   |      |      |      |      |      |    |
| 20. laziness       | -0.08 | -0.16 | -0.1  | -0.13 | -0.01 | -0.15 | -0.08 | -0.19 | -0.1  | -0.05 | -0.01 | -0.08 | 0.12 | 0.31  | -0.17 | 0.05  | -0.11 | 0.22 | 0.2  | --   |      |      |      |      |    |
| 21. pride          | -0.01 | 0.05  | -0.04 | 0.0   | 0.0   | 0.03  | -0.06 | -0.05 | 0.05  | -0.08 | 0.21  | 0.07  | 0.19 | 0.1   | 0.03  | -0.05 | 0.03  | 0.28 | 0.13 | 0.01 | --   |      |      |      |    |
| 22. spite          | -0.2  | -0.19 | -0.14 | -0.08 | -0.05 | -0.15 | -0.1  | -0.09 | -0.1  | -0.07 | 0.07  | -0.08 | 0.03 | 0.24  | -0.22 | -0.03 | -0.1  | 0.21 | 0.14 | 0.27 | 0.32 | --   |      |      |    |
| 23. jealousy       | -0.25 | -0.27 | -0.14 | -0.18 | -0.08 | -0.24 | -0.17 | -0.18 | -0.12 | -0.08 | 0.04  | -0.05 | 0.06 | 0.25  | -0.27 | 0.01  | -0.15 | 0.21 | 0.22 | 0.54 | 0.12 | 0.61 | --   |      |    |
| 24. competition    | 0.02  | -0.12 | -0.06 | 0.02  | 0.04  | 0.0   | -0.05 | -0.07 | 0.11  | 0.04  | 0.19  | -0.01 | 0.2  | 0.26  | -0.07 | -0.03 | 0.0   | 0.16 | 0.35 | 0.24 | 0.21 | 0.37 | 0.3  | --   |    |
| 25. revenge        | -0.1  | -0.17 | -0.08 | -0.05 | -0.02 | -0.02 | -0.11 | -0.26 | -0.03 | -0.19 | 0.0   | -0.01 | 0.14 | 0.27  | -0.09 | -0.05 | -0.08 | 0.38 | 0.13 | 0.26 | 0.16 | 0.32 | 0.37 | 0.36 | -- |

*Note.* This table displays the Pearson correlation coefficients between traits (items 1–10) and motives (items 11–25) based on participants' meta-perceptions of how their cooperative behaviours might be observed by others.

**Table C4.2***Correlation Matrix of Traits and Motives for Uncooperative Behaviours*

|                    | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    | 11    | 12    | 13    | 14    | 15    | 16    | 17    | 18    | 19    | 20    | 21   | 22   | 23   | 24   | 25 |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|------|------|------|----|
| 1. trustworthy     | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 2. sincere         | 0.69  | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 3. honest          | 0.57  | 0.65  | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 4. loyal           | 0.75  | 0.64  | 0.56  | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 5. just            | 0.58  | 0.52  | 0.53  | 0.51  | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 6. reliable        | 0.78  | 0.61  | 0.53  | 0.69  | 0.57  | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 7. empathic        | 0.69  | 0.64  | 0.5   | 0.63  | 0.52  | 0.65  | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 8. likeable        | 0.66  | 0.55  | 0.55  | 0.62  | 0.55  | 0.65  | 0.72  | --    |       |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 9. competent       | 0.69  | 0.55  | 0.45  | 0.55  | 0.6   | 0.63  | 0.49  | 0.62  | --    |       |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 10. intelligent    | 0.52  | 0.43  | 0.52  | 0.39  | 0.5   | 0.52  | 0.44  | 0.54  | 0.65  | --    |       |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 11. reputation     | 0.26  | 0.17  | 0.25  | 0.3   | 0.16  | 0.24  | 0.15  | 0.16  | 0.17  | 0.09  | --    |       |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 12. selflessness   | 0.2   | 0.11  | 0.08  | 0.15  | 0.17  | 0.15  | 0.16  | 0.11  | 0.07  | 0.1   | 0.16  | --    |       |       |       |       |       |       |       |       |      |      |      |      |    |
| 13. obligation     | 0.28  | 0.25  | 0.19  | 0.24  | 0.31  | 0.22  | 0.21  | 0.21  | 0.27  | 0.23  | 0.22  | 0.24  | --    |       |       |       |       |       |       |       |      |      |      |      |    |
| 14. avoid conflict | 0.26  | 0.19  | 0.06  | 0.25  | 0.06  | 0.26  | 0.31  | 0.34  | 0.23  | 0.22  | 0.28  | 0.14  | 0.25  | --    |       |       |       |       |       |       |      |      |      |      |    |
| 15. helping        | 0.48  | 0.4   | 0.32  | 0.4   | 0.43  | 0.39  | 0.48  | 0.44  | 0.36  | 0.25  | 0.21  | 0.32  | 0.38  | 0.26  | --    |       |       |       |       |       |      |      |      |      |    |
| 16. fairness       | 0.5   | 0.46  | 0.39  | 0.43  | 0.59  | 0.49  | 0.44  | 0.49  | 0.51  | 0.37  | 0.28  | 0.28  | 0.39  | 0.24  | 0.41  | --    |       |       |       |       |      |      |      |      |    |
| 17. empathy        | 0.46  | 0.39  | 0.31  | 0.37  | 0.41  | 0.43  | 0.52  | 0.47  | 0.31  | 0.21  | 0.34  | 0.33  | 0.38  | 0.35  | 0.76  | 0.46  | --    |       |       |       |      |      |      |      |    |
| 18. guilt          | 0.18  | 0.14  | 0.1   | 0.16  | 0.22  | 0.23  | 0.23  | 0.27  | 0.23  | 0.2   | 0.28  | 0.25  | 0.37  | 0.48  | 0.31  | 0.29  | 0.36  | --    |       |       |      |      |      |      |    |
| 19. selfishness    | -0.19 | -0.12 | -0.11 | -0.12 | -0.21 | -0.15 | -0.28 | -0.22 | -0.01 | -0.04 | -0.05 | -0.19 | 0.01  | -0.07 | -0.36 | -0.1  | -0.35 | -0.1  | --    |       |      |      |      |      |    |
| 20. laziness       | -0.28 | -0.3  | -0.28 | -0.29 | -0.38 | -0.4  | -0.22 | -0.25 | -0.32 | -0.27 | -0.17 | -0.05 | -0.18 | -0.14 | -0.17 | -0.41 | -0.18 | -0.13 | 0.12  | --    |      |      |      |      |    |
| 21. pride          | 0.02  | 0.05  | 0.11  | 0.13  | 0.12  | 0.06  | -0.04 | 0.05  | 0.12  | 0.07  | 0.28  | 0.05  | 0.15  | 0.16  | 0.11  | 0.23  | 0.12  | 0.21  | 0.06  | -0.17 | --   |      |      |      |    |
| 22. spite          | -0.19 | -0.15 | -0.1  | -0.09 | 0.07  | -0.05 | -0.15 | -0.08 | 0.0   | 0.04  | 0.13  | 0.12  | 0.13  | 0.0   | 0.0   | 0.19  | 0.07  | 0.2   | 0.04  | -0.03 | 0.4  | --   |      |      |    |
| 23. jealousy       | -0.16 | -0.17 | -0.15 | -0.16 | -0.09 | -0.15 | -0.18 | -0.25 | -0.17 | -0.1  | 0.23  | 0.09  | 0.04  | -0.01 | -0.06 | 0.03  | 0.02  | 0.2   | 0.04  | 0.05  | 0.34 | 0.52 | --   |      |    |
| 24. competition    | 0.03  | 0.02  | 0.05  | 0.08  | 0.13  | 0.08  | -0.1  | -0.06 | 0.11  | 0.11  | 0.3   | 0.22  | 0.24  | 0.05  | 0.08  | 0.15  | 0.13  | 0.25  | 0.1   | -0.17 | 0.5  | 0.39 | 0.5  | --   |    |
| 25. revenge        | -0.14 | -0.19 | -0.15 | -0.1  | 0.09  | -0.05 | -0.16 | -0.12 | 0.0   | 0.06  | 0.09  | 0.14  | 0.09  | 0.03  | -0.02 | 0.22  | 0.06  | 0.08  | -0.03 | -0.14 | 0.38 | 0.78 | 0.54 | 0.44 | -- |

*Note.* This table displays the Pearson correlation coefficients between traits (items 1–10) and motives (items 11–25) based on participants' meta-perceptions of how their uncooperative behaviours might be observed by others.



## C5 A Sample of Free Text Responses

This section presents a selection of participant responses to illustrate the variety of cooperative and uncooperative behaviours they reported. These responses encompass a wide range of behaviours observed in different situations, involving diverse relationships among participants and others, described at varying levels of detail, and written with differing focuses (e.g., explaining situations, justifying behaviours, or highlighting motivations).

### *Cooperative Behaviours*

Participants described a variety of cooperative actions, including giving thoughtful gifts (*“I know one of my coworkers likes the laughing cow cheese a lot. I have recently went to the shops and got some of those cheese to snack on. Knowing that she likes the cheese, I brought her a piece when I went into the office”*), making donations when prompted (*“When I went to a fast food outlet. I was asked if I want to donate a certain amount to help feed less fortunate children. I agreed to donate”*), and following rules rather than defecting (*“I was boarding a bus in Germany. I bought a bus ticket as i should when boarding a bus”*).

These cooperative actions took place in various contexts, such as at work (*“A team member was unable to fulfill their shift because of personal reasons, so I voluntarily decided to take up their shift as well”*), at home (*“Last week, I noticed that one of my roommates was completely busy studying for many final exams. She was only eating junk food and drinking a lot of coffee and energy drinks. In order to help her, I cooked a nice healthy meal for her and leaved by his room door”*), and at the gym (*“I regularly climb in a local gym, and reached intermediate level. Yesterday a guy started trying a route I’ve already climbed, right next to the one I was working on. I saw he had issues starting off so I gave some tips and helped him find a better way up”*).

Many examples of cooperative behaviour benefited both specific strangers (*“I helped a lady jumpstart her car. Her car wouldn’t start and she was stopped at a gas station”*) and unspecific strangers (*“I had decided to empty my cupboards of unused items, which I gave to a charity shop. this included clothes, books and ornaments”*).

Some participants provided caveats to their cooperative behaviours (*“In my job I always prepare customers documents as soon as possible. Only exception is if they are impolite”*), highlighted personal costs and the targets’ appreciation (*“My partner was struggling with several bags of groceries. Although I was tired from a long day at work and it was cold outside, I went over to lend a hand. Together, we brought in the groceries quickly and easily. It was a*

*simple act of cooperation, but my partner appreciated the support, and appreciated me noticing that he could have used a helping hand”), or explanations for doing a potentially bad thing out of good motives (“I know this is wrong, but I need you to read everything before you judge me. I recently entered university, and within it I made quite a few friends. I have a specific friend with whom I am very close since I go home with him, plus we are part of the same group of friends. In the second semester of last year we had a physics class. At my school I had previously chosen a physics specialty, so I knew a lot about the subject and I was doing well. When we had our last evaluation of the branch, my friend asked me for help, he asked me if he could copy me in that evaluation, this since he was about to repeat that branch. I completely agreed since I have no problem helping a friend when they need it, even more so if that friend has been faithful for a long time, he is a real friend. It should be noted that the field of physics has nothing to do with our university career, since we are from the health area”).*

The same cooperative action, such as giving a lift, was performed with varying levels of cost and relationship closeness: from picking up teammates (*“i picked up teammates from my chess team. We played together so i picked them up with the car”*), to a more inconvenient (one-time) lift for a friend (*“Recently a friend of mine told me they had a flight going out from my city at 5 am and needed to sleep in my home and asked me if I could drive them to the airport, which I did”*), to a regularly costly act for a romantic partner (*“I give me wife lifts to work all the time 2 hour ride there and 2 hour back. I do this because she works hard”*). The latter participant also provided a reason for why he does so (she works hard), showing that he is reciprocating the efforts of his partner.

Participants also emphasised different aspects when describing blood donation, such as traits of dedication and reliability (*“i donate blood every 3 months without failure. I have never missed an appointment”*), or a motivation of repayment (*“Once many years ago i had a serious surgery and i relied to others for the necessary blood donations. So since then and every three months i visit the local hospital to donate blood, in an effort to repay my debt”*).

Some descriptions centred around collective efforts that required varying levels of effort from the participant, such as participating in a clean-up initiative (*“I took part in a clean-up initiative in my neighborhood. I alongside a dozen others worked together to clean up litter and tidy up the streets”*), or voting for a communal action to be carried out (*“We as a community wanted to upgrade a dirt road. The whole street had to pitch in and help with filling hokes and remove boulders. I voted yes that we as a street should do it”*).

## ***Uncooperative Behaviours***

Participants described a variety of uncooperative actions, such as missing an opportunity to provide (unrequested) assistance (*"Some time ago, while I was walking home I saw an old lady who was carrying her groceries. I was walking behind her and I thought about helping her because her bag looked heavy but I didn't. I walked past her and went home"*), engaging in petty punishment (*"When I argue with my sibling I become very petty. A few days ago I had made cookies for me and my family. The next day I had a small argument with my younger brother because I felt he was being selfish; I hid the cookies from him, but kept offering them to the rest of my family, because I wanted to be selfish back at him"*), inconveniencing another to continue talking to a friend (*"I was approached by a customer who wanted to get booked into their appointment. I was talking to a work friend and instead of booking them in, which i could've done in less than a minute, i sent them back to the front of the store where there is a small queue for you to be checked in"*), and acting competitively out of fear (*"In my work environment, I have noticed that sometimes I tend to withhold information, not sharing it with my colleagues, primarily due to an underlying fear that by doing so, I might enhance their productivity and, consequently, be surpassed in the speed with which they carry out their tasks"*).

Many participants explained that their uncooperative actions were a result of others' lack of cooperation, such as being tired of picking up after others (*"I didn't clean my home because my brothers kept messing it up. I was tired of picking up after adults"*), refusing to cooperate with uncooperative project partners (*"Last semester we had project to do. People from my group were annoying and they did not want to cooperate. So I thought why would I cooperate with them? Then started working on my own, not responding to texts they were sending me"*), or slacking due to not receiving an earned wage increase (*"I have been working for a company in an administrative position for 5 years and have not received my wage increase which was rightfully due to me, despite my many requests for my increase, I still did not get an answer so I began slacking at my job and not doing my duties timeously and/or sometimes at all"*).

Participants frequently explained why they acted uncooperatively, citing reasons such as being fed up and disliking the other person (*"At work I denied to help my coworker with his task even though I was not very busy with my own tasks. I was tired and not in a good mood that day. And to be honest I dislike this person for over-using the help of other people. Thats why I denied to help and said I was too busy to help"*), not thinking help was deserved (*"i passed a beggar next to the road and refuse to give him anything as he was not disabled so could go and find*

*work instead of begging”), needing to prioritise their own work over assisting others (“I work some shifts on a reception and often get interrupted by people knocking on the door or window to be let in. This is really annoying when I am in the middle of something. This week somebody stood in the window waving at me to let them in, but I was in the middle of a task and was already in a bad mood, so when I saw them in my peripheral vision I carried on with my work and made sure not to look their way, so I could pretend I hadn't seen them”), being tired (“I would usually do anything for anyone but on a recent occasion when I was asked to help in the local community shop I lied and said I was busy when really I was just wanting a bit of time to myself. I was washed out and needed some me time”), being tired and there being available alternatives (“My friend asked me to pick him up at the airport and take him home, but I was tired and really didn't want to drive till there, and since there is a bus system near the airport I told him I couldn't do it”), and providing mitigating explanations such as not having a poo bag (“I was walking my dog and we were passing through a suburban area. My dog pooped on a lawn (that was a common area, not private property) and I did not have any poo bags with me, so I left without cleaning up after him”).*

Many participants were also upfront about being selfish, without providing excuses, and often included the impact of their selfishness on others. Examples include steamrolling others (“During a recent group outing with friends, I behaved uncooperatively by not considering everyone's preferences when planning activities. Instead, I insisted on my choices without considering the interests of others, leading to some feeling left out and dissatisfied”), making a delivery man wait (“On a Friday night, I was watching a movie and decided to order a burger. The delivery guy arrived at the same time of the last minutes of the movie, so I intentionally delayed to collect it in order to finish the movie”), not being a team player in sports (“In the last game on the weekend (I play rugby), we were losing by 3 points which was a very close score. As the match advanced I made a call to do one incorrect play, but I wanted to try myself. I ignored what another teammate told me and picked up the ball and lost it at the play. There was a clear opportunity to score outside but I wanted to do the try myself, and in consequence, we couldn't win the match”), and leaving the laundry on the washing line (“I was too lazy to remove my laundry from the washing line last night - meaning other people would have to wait on me to do so before they could hang theirs”).

# Appendix D

## Supplementary Information for Chapter 5

### D1 Preregistration for “Investigating Observer Perceptions of Motives in Second- and Third-Party Punishers through Vignettes”

#### Motivations of third-party and second-party punishers (#134404)

##### Author(s)

**Pre-registered on:** 06/02/2023 03:02 PM (PT)

This pre-registration is currently anonymous to enable blind peer-review.  
 It has one author.

##### 1) Have any data been collected for this study already?

No, no data have been collected for this study yet.

##### 2) What's the main question being asked or hypothesis being tested in this study?

- 1) Punishing actions will generally reflect positively on punishers.
- 2) Third-party punishers will be evaluated more positively than second-party punishers.
- 3) Prosocial motive items will be more relevant for third-party punishment than second-party punishment, and formidability motive items will be morerelevant to second-party punishment than third-party punishment.

##### 3) Describe the key dependent variable(s) specifying how they will be measured.

- 18 motive items from 1 = "Not at all" to 7 = "Completely"
- overall reflection of punishing actions on the punisher from 1 = "Extremely negatively" to 7 = "Extremely positively"

##### 4) How many and which conditions will participants be assigned to?

Participants will not be split into conditions per se, however there are seven scenarios for which exists both a third-party punishment and a second-party punishment scenario. Participants will be randomly presented with one of these versions (the random presentation of TPP vs SPP occurs for each scenario anew, TPP vs SPP items are evenly presented).

##### 5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

- 1) I will compare participants punisher evaluation scores to the scale midpoint (4) to detect below or above average effects. For this, participants evaluationscores will be subtracted by four, which will then be used as the dependent variable in an intercept only model with participant ID as random effect ( $\text{adjusted\_evaluation\_score} \sim 1 + (1 | \text{ID})$ ).
- 2) I will conduct a linear regression with punisher evaluation score as dependent variable, punishing type (1 = TPP, 0 = SPP) as fixed effect and participant IDas random effect ( $\text{evaluation\_score} \sim \text{type} + (1 | \text{ID})$ ).

3) I will conduct a linear regression with motive score as dependent variable and item (18 levels) as well as punishing type as fixed effects, with an interaction between them. Participants will again be a random effect ( $\text{motive\_score} \sim \text{type} * \text{item} + (1 | \text{ID})$ ).

**6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.**

We will exclude participants who give the same response across likert scales and participants who fail the attention check question.

**7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.**

I will post the study in the subreddit r/SampleSize and will repost it for two to four weeks.

**8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)** N/A

**Available at** <https://aspredicted.org/c9cp-446h.pdf>

Version of AsPredicted Questions: 2.00

## D2 Study Materials for the Vignette Study

---

### Start of Block: Info and consent

info *This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani in the Department of Experimental Psychology at UCL.

**In this study, you will read scenarios in which someone responds to another person's wrong-doing. We will ask you to indicate what you think motivated the individual's response.**

The study is estimated to last around 5-10 minutes.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. Only researchers working with the Raihani Lab will have access to datafiles containing identifying information. Completely de-identified datafiles might be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please close the study or select "I do not consent to take part in this study".

---



consent Please only consent to taking part in this study if you agree with ALL of the following statements:

(1) I have read the information above and understand I can email the researchers with any questions. (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing. (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications. (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared). (5) I understand that I am free to withdraw from the study without penalty if I so wish.

- ☐ I **consent** to take part in this study. (1)
- ☐ I do **NOT consent** to take part in this study. (0)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent </strong>to take part in this study.*

End of Block: Info and consent

---

Start of Block: Demographics



age What is your age in years?

---

gender What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Non-binary / third gender (3)
- ☐ Prefer not to say (4)

End of Block: Demographics

Start of Block: TPP1

T1S At the gym, Sarah witnesses Mark consistently disrupting another gym member's workout. Sarah therefore reports Mark's behaviour to the gym, which revokes Mark's membership.

T1M To what extent do you think Sarah was motivated by the following:

|                        | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| appear cooperative (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| be helpful (2)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| show strength (3)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| signal dominance (4)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| assert authority (5)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seek revenge (6)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| protect herself (7)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |



|                                           |                       |                       |                       |                       |                       |                       |                       |
|-------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| protect someone else (8)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| do the right thing (9)                    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| enforce rules or regulations (10)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| maintain social order (11)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| deter future misconduct (12)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| express disapproval (13)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seek justice (14)                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| restore balance or fairness (15)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| teach a lesson (16)                       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| defend personal or collective values (17) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| gain respect or admiration (18)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

T1R Do you think punishing Mark reflected more negatively or more positively on Sarah?

|     | Extremely negatively (1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely positively (7) |
|-----|--------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------|
| (1) | <input type="radio"/>    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>    |

End of Block: TPP1

### Start of Block: SPP1

S1S At the gym, Sarah's workout is consistently disrupted by Mark, another gym member. Sarah therefore reports Mark's behaviour to the gym, which revokes Mark's membership.



S1M To what extent do you think Sarah was **motivated** by the following:  
[same 18-item motivation table as above]

S1R Do you think punishing Mark reflected more negatively or more positively on Sarah?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

### End of Block: SPP1

### Start of Block: TPP2

T2S As Jessica walks along the street she witnesses a mugging. However, Jessica rushes over, yanks the handbag out of the would-be muggers hands and, as he starts running away, hits him with it.



T2M To what extent do you think Jessica was **motivated** by the following:  
[same 18-item motivation table as above]

T2R Do you think punishing the mugger reflected more negatively or more positively on Jessica?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: TPP2

Start of Block: SPP2

S2S As Jessica walks along the street she is mugged. However, Jessica yanks the handbag out of the would-be muggers hands and, as he starts running away, hits him with it.



S2M To what extent do you think Jessica was **motivated** by the following:

*[same 18-item motivation table as above]*

S2R Do you think punishing the mugger reflected more negatively or more positively on Jessica?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: SPP2

Start of Block: TPP3

T3S At a community gathering, Dave witnesses a person repeatedly making derogatory and offensive comments towards certain individuals. In response, Dave decides to exclude the person from future community events and gatherings.



T3M To what extent do you think Dave was **motivated** by the following:  
[same 18-item motivation table as above]

T3R Do you think punishing the person reflected more negatively or more positively on Dave?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: TPP3

Start of Block: SPP3

S3S At a community gathering a person repeatedly makes derogatory and offensive comments towards Dave. In response, Dave decides to exclude the person from future community events and gatherings.



S3M To what extent do you think Dave was **motivated** by the following:  
[same 18-item motivation table as above]

S3R Do you think punishing the person reflected more negatively or more positively on Dave?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: SPP3

Start of Block: TPP4

T4S Alex witnesses someone damaging his parent’s property and makes them compensate his parents for the damages.



T4M To what extent do you think Alex was **motivated** by the following:  
*[same 18-item motivation table as above]*

T4R Do you think punishing the person reflected more negatively or more positively on Alex?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: TPP4

Start of Block: SPP4

S4S Alex witnesses someone damaging his property and makes them compensate him for the damages.



S4M To what extent do you think Alex was **motivated** by the following:  
[same 18-item motivation table as above]

---

S4R Do you think punishing the person reflected more negatively or more positively on Alex?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: SPP4

---

Start of Block: TPP5

T5S Emma becomes aware of a co-worker consistently taking credit for other people's achievements. So, Emma discreetly shares this information with other colleagues, leading to a decline in the other person's professional reputation and credibility.



T5M To what extent do you think Emma was **motivated** by the following:  
[same 18-item motivation table as above]

---

T5R Do you think punishing the person reflected more negatively or more positively on Emma?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: TPP5

---

Start of Block: SPP5

S5S Emma becomes aware of a co-worker consistently taking credit for her achievements. So, Emma discreetly shares this information with other colleagues, leading to a decline in the other person's professional reputation and credibility.



S5M To what extent do you think Emma was **motivated** by the following:

*[same 18-item motivation table as above]*

S5R Do you think punishing the person reflected more negatively or more positively on Emma?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: SPP5

Start of Block: TPP6

T6S John discovers that Kim stole ten pounds from their friend and decides to stop giving her rides, which he usually does.



T6M To what extent do you think John was **motivated** by the following:

*[same 18-item motivation table as above]*

T6R Do you think punishing Kim reflected more negatively or more positively on John?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: TPP6

Start of Block: SPP6

S6S John discovers that Kim stole ten pounds from him and decides to stop giving her rides, which he usually does.

S6M To what extent do you think John was **motivated** by the following:  
[same 18-item motivation table as above]

S6R Do you think punishing Kim reflected more negatively or more positively on John?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: SPP6

Start of Block: TPP7

T7S Mary and Jim are sharing an extra large pepperoni pizza while Tamara has her own small vegetarian pizza. When Jim eats Tamara's last slice without asking, Mary tells him off for being inconsiderate and selfish.





T7M To what extent do you think Mary was **motivated** by the following:  
[same 18-item motivation table as above]

T7R Do you think punishing Jim reflected more negatively or more positively on Mary?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: TPP7

Start of Block: SPP7

S7S Mary and Jim are sharing an extra large pepperoni pizza while Tamara has her own small vegetarian pizza. When Jim eats Tamara's last slice without asking, she tells him off for being inconsiderate and selfish.



S7M To what extent do you think Tamara was **motivated** by the following:  
[same 18-item motivation table as above]

S7R Do you think punishing Jim reflected more negatively or more positively on Tamara?

|     | Extremely<br>negatively<br>(1) | (2)                   | (3)                   | Neither (4)           | (5)                   | (6)                   | Extremely<br>positively<br>(7) |
|-----|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------------|
| (1) | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          |

End of Block: SPP7

Start of Block: attention check

attention Attention check:

What have the scenarios been about?

(please do not respond incorrectly for humorous reasons as your responses will have to be excluded)

- ☐ consumer purchasing habits (1)
- ☐ responses to wrong-doings (2)
- ☐ popular sports (3)
- ☐ book preferences (4)

End of Block: attention check

---

Start of Block: Debrief

### **Debrief**

*Thank you for taking part in this study!*

Each of the seven scenarios you were presented with had two version: one in which the punisher was the victim ("second-party punishment"), and one in which someone other than the punisher was the victim ("third-party punishment").

With your responses I will investigate: whether punishers were generally evaluated more positively or negatively whether punishers received more favourable evaluations when they themselves had not been the victim but were instead stepping in for someone else which motive items were consistently more relevant for second-party vs. third-party punishment. For example, we expect third-party punishers to be attributed more prosocial motives, such as wanting to help, whereas we expect second-party punishers to be attributed motives related to formidability, such as wanting to appear strong.

*Optional Comment Box:*

---

End of Block: Debrief

---

### **D3 Evaluations of Punishers Based on Punishing Type: Analysis Without Controlling for Story ID**

To investigate whether third-party punishers receive more positive evaluations compared to second-party punishers, a linear mixed effects model was fit. Participants' evaluations of punishers were used as the response variable, while punishing type (1 = TPP, 0 = SPP) was included as a fixed effect, and participant ID was specified as the random effect. As preregistered, this version does not include Story ID as a control. Reflecting findings from the 5.1.3 results section, third-party punishers were perceived significantly more positively than second-party punishers,  $b = 0.2$ , 95% CI [0.15, 0.25],  $t(10159.47) = 7.98$ ,  $p < .0001$ .

## D4 Correlation Matrices of Motives Attributed to Second and Third-Party Punishers

**Table D4.1**

*Correlation Matrix of Motives for SPP Behaviours*

|                         | 1     | 2     | 3    | 4    | 5    | 6     | 7     | 8     | 9    | 10   | 11   | 12   | 13    | 14   | 15   | 16   | 17   | 18 |
|-------------------------|-------|-------|------|------|------|-------|-------|-------|------|------|------|------|-------|------|------|------|------|----|
| 1. appear cooperative   | --    |       |      |      |      |       |       |       |      |      |      |      |       |      |      |      |      |    |
| 2. be helpful           | 0.52  | --    |      |      |      |       |       |       |      |      |      |      |       |      |      |      |      |    |
| 3. show strength        | 0.09  | 0.1   | --   |      |      |       |       |       |      |      |      |      |       |      |      |      |      |    |
| 4. signal dominance     | 0.19  | 0.24  | 0.59 | --   |      |       |       |       |      |      |      |      |       |      |      |      |      |    |
| 5. assert authority     | 0.22  | 0.24  | 0.52 | 0.65 | --   |       |       |       |      |      |      |      |       |      |      |      |      |    |
| 6. seek revenge         | 0.12  | 0.12  | 0.28 | 0.47 | 0.41 | --    |       |       |      |      |      |      |       |      |      |      |      |    |
| 7. protect self         | -0.06 | -0.01 | 0.16 | 0.11 | 0.2  | 0.15  | --    |       |      |      |      |      |       |      |      |      |      |    |
| 8. protect another      | 0.37  | 0.58  | 0.05 | 0.12 | 0.13 | 0.01  | 0.05  | --    |      |      |      |      |       |      |      |      |      |    |
| 9. do the right thing   | 0.3   | 0.41  | 0.14 | 0.2  | 0.33 | 0.08  | 0.19  | 0.28  | --   |      |      |      |       |      |      |      |      |    |
| 10. enforce rules       | 0.15  | 0.21  | 0.22 | 0.24 | 0.33 | 0.11  | 0.23  | 0.18  | 0.44 | --   |      |      |       |      |      |      |      |    |
| 11. maintain order      | 0.23  | 0.21  | 0.22 | 0.19 | 0.25 | -0.02 | 0.1   | 0.14  | 0.26 | 0.43 | --   |      |       |      |      |      |      |    |
| 12. deter misconduct    | 0     | -0.02 | 0.13 | 0.01 | 0.09 | -0.08 | 0.19  | 0.06  | 0.2  | 0.3  | 0.19 | --   |       |      |      |      |      |    |
| 13. express disapproval | -0.12 | -0.08 | 0.1  | 0.05 | 0.09 | -0.01 | 0.11  | -0.08 | 0.11 | 0.06 | 0.05 | 0.33 | --    |      |      |      |      |    |
| 14. seek justice        | 0.15  | 0.02  | 0.25 | 0.19 | 0.26 | 0.3   | 0.39  | 0.02  | 0.26 | 0.32 | 0.11 | 0.08 | 0.08  | --   |      |      |      |    |
| 15. restore fairness    | 0.14  | 0.11  | 0.06 | 0.02 | 0.08 | 0.05  | 0.31  | 0.11  | 0.34 | 0.27 | 0.21 | 0.14 | 0.13  | 0.52 | --   |      |      |    |
| 16. teach a lesson      | 0.09  | 0.11  | 0.27 | 0.31 | 0.31 | 0.4   | 0.17  | 0.07  | 0.11 | 0.17 | 0.11 | 0.22 | 0.23  | 0.32 | 0.2  | --   |      |    |
| 17. defend values       | 0.12  | 0.15  | 0.14 | 0.07 | 0.29 | 0.03  | 0.24  | 0.12  | 0.49 | 0.32 | 0.35 | 0.29 | 0.33  | 0.17 | 0.22 | 0.12 | --   |    |
| 18. gain respect        | 0.36  | 0.24  | 0.17 | 0.17 | 0.07 | 0.08  | -0.02 | 0.14  | 0.02 | 0.06 | 0.15 | 0.01 | -0.04 | 0.12 | 0.12 | 0.1  | 0.03 | -- |

*Note.* This table displays the Pearson correlation coefficients between motive items in the context of second-party punishment.

**Table D4.2***Correlation Matrix of Motives for TPP Behaviours*

|                         | 1    | 2     | 3    | 4     | 5     | 6     | 7     | 8     | 9     | 10   | 11   | 12    | 13   | 14   | 15   | 16   | 17   | 18 |
|-------------------------|------|-------|------|-------|-------|-------|-------|-------|-------|------|------|-------|------|------|------|------|------|----|
| 1. appear cooperative   | --   |       |      |       |       |       |       |       |       |      |      |       |      |      |      |      |      |    |
| 2. be helpful           | 0.35 | --    |      |       |       |       |       |       |       |      |      |       |      |      |      |      |      |    |
| 3. show strength        | 0.11 | -0.03 | --   |       |       |       |       |       |       |      |      |       |      |      |      |      |      |    |
| 4. signal dominance     | 0.15 | -0.02 | 0.66 | --    |       |       |       |       |       |      |      |       |      |      |      |      |      |    |
| 5. assert authority     | 0.23 | -0.03 | 0.68 | 0.72  | --    |       |       |       |       |      |      |       |      |      |      |      |      |    |
| 6. seek revenge         | -0.1 | -0.12 | 0.35 | 0.46  | 0.43  | --    |       |       |       |      |      |       |      |      |      |      |      |    |
| 7. protect self         | 0.04 | -0.07 | 0.33 | 0.27  | 0.27  | 0.19  | --    |       |       |      |      |       |      |      |      |      |      |    |
| 8. protect another      | 0.23 | 0.58  | 0.01 | -0.06 | -0.04 | -0.12 | -0.11 | --    |       |      |      |       |      |      |      |      |      |    |
| 9. do the right thing   | 0.14 | 0.5   | 0    | -0.11 | -0.08 | -0.22 | -0.03 | 0.54  | --    |      |      |       |      |      |      |      |      |    |
| 10. enforce rules       | 0.16 | 0.31  | 0.28 | 0.27  | 0.3   | 0.12  | 0.14  | 0.33  | 0.36  | --   |      |       |      |      |      |      |      |    |
| 11. maintain order      | 0.2  | 0.36  | 0.06 | 0.02  | 0.07  | -0.11 | -0.03 | 0.34  | 0.39  | 0.44 | --   |       |      |      |      |      |      |    |
| 12. deter misconduct    | 0.04 | 0.18  | 0    | -0.11 | -0.06 | -0.1  | 0.08  | 0.24  | 0.34  | 0.28 | 0.31 | --    |      |      |      |      |      |    |
| 13. express disapproval | 0.12 | 0.11  | 0.04 | 0.05  | 0.07  | -0.04 | 0.01  | 0.12  | 0.2   | 0.18 | 0.24 | 0.27  | --   |      |      |      |      |    |
| 14. seek justice        | 0.08 | 0.29  | 0.14 | 0.16  | 0.13  | 0.11  | 0.05  | 0.39  | 0.27  | 0.22 | 0.31 | 0.17  | 0.07 | --   |      |      |      |    |
| 15. restore fairness    | 0.12 | 0.34  | 0.02 | 0.01  | 0.03  | 0     | 0.09  | 0.38  | 0.38  | 0.3  | 0.37 | 0.28  | 0.22 | 0.53 | --   |      |      |    |
| 16. teach a lesson      | 0.1  | 0.01  | 0.32 | 0.34  | 0.35  | 0.28  | 0     | 0.04  | 0.08  | 0.2  | 0.18 | 0.22  | 0.3  | 0.28 | 0.13 | --   |      |    |
| 17. defend values       | 0.21 | 0.28  | 0.1  | -0.06 | 0.03  | -0.18 | 0.02  | 0.29  | 0.47  | 0.29 | 0.39 | 0.32  | 0.42 | 0.18 | 0.34 | 0.03 | --   |    |
| 18. gain respect        | 0.24 | 0.02  | 0.55 | 0.62  | 0.63  | 0.35  | 0.25  | -0.03 | -0.04 | 0.18 | 0.05 | -0.06 | 0.01 | 0.08 | 0.07 | 0.19 | 0.01 | -- |

*Note.* This table displays the Pearson correlation coefficients between motive items in the context of third-party punishment.

## D5 Preregistration for “Meta-Cognitive and Observer Accuracy Surrounding Punisher Motive Attributions”

### Observer- and meta-accuracy for second- and third-party punishment (#162739)

#### Author(s)

Pre-registered on: 02/19/2024 09:14 AM (PT)

This pre-registration is currently anonymous to enable blind peer-review.  
 It has one author.

#### 1) Have any data been collected for this study already?

No, no data have been collected for this study yet.

#### 2) What's the main question being asked or hypothesis being tested in this study?

We will investigate whether people know how their punishing actions are viewed by others (meta-accuracy) and whether observers' perceptions match punishers' actual reported motives and traits (observer accuracy). Our main predictions are, that:

- participants will exhibit observer- and meta-accuracy
- whenever punishers do misestimate how they are perceived, they will believe others would judge them less favourably than they actually do, indicating a systematic negative meta-bias

#### 3) Describe the key dependent variable(s) specifying how they will be measured.

We will solicit descriptions of a second-party punishment (SPP) and a third-party punishment (TPP) behaviour participants recently engaged in, with accompanying self- and meta-perceptions. Observer participants will then provide actual judgements on a random sub-set of those behaviour descriptions.

Self-perceptions, meta-perceptions, and actual judgements will be given for the same 23 motive and 7 trait items, all rated from 1 = "Not at all" to 7 = "Completely" (i.e., 30 ratings each for meta-perceptions, self-perceptions and actual judgments per behaviour). Estimates provided by participants (i.e. meta-perceptions for meta-accuracy and judgements for observer accuracy) will be modelled as the response variables, and the true values (i.e. observer judgements for meta-accuracy and punishers' self-perceptions for observer accuracy) will be modelled as predictors. We will not run analyses directly with Likert-scale responses, and will instead use cluster centring within actor to ensure that perceptions are anchored to the mean validation criterion within each actor (i.e. punisher) participant (for more detail see end of Question 5).

#### 4) How many and which conditions will participants be assigned to?

Data collection will occur in two stages: first we will recruit actor participants who will each provide one description of a recent TPP and one description of a recent SPP behaviour they engaged in, with accompanying self- and meta-perceptions. In a second stage, observers will read a random subset of these behaviour descriptions (4 TPP and 4 SPP behaviours) and provide judgements.

#### 5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

We will fit three main mixed-effects models when investigating meta-accuracy and observer accuracy. Models will be fit separately for TPP and SPP behaviours.

Baseline accuracy models test for a linear relationship between estimated and true values in actor meta-perceptions (meta-perception ~ observer judgment) and observer attributions

(observer judgment ~ self-perception). Distinctive accuracy model test the relationship between estimated values and the normative profile on top of the baseline models' relationship between the estimated and true values. The normative profile is the grand mean centred distribution of the self-perception means across actors. The third models test for insight. Actor insight is the relationship between meta-perceptions and judgements while controlling for self-perceptions. As the model controls for self-perception, the relationship between meta-perceptions and actual judgements can be interpreted as actors' understanding of how they will be misperceived. The relationship between meta-perception and self-perception shows transparency bias – the extent to which actors overestimate how accurately they will be perceived. Observer insight is the relationship between judgements and self-perceptions while controlling for meta-perceptions.

To examine systematic biases in meta-perceptions we will variable-centre the data and interact item with observer judgment in the baseline meta-accuracy model (meta-perceptions centred on the mean observer judgment of each item and observers' judgments on the mean self-perception of each item).

Values of zero represent average point estimate accuracy by item (i.e., estimates which differ from zero show systematic directional bias).

We will adopted a maximal random structures approach, with random slopes and intercepts for all linear predictors across all random effects. Whenever a singular fit is obtained we will remove random slopes in the order of interactions, moderating variables, confounding variables (e.g., normative profile) and true values. When predicting actor meta-accuracy a single random effect (actor) will be specified, and when predicting observer accuracy two crossed random effects will be included (actor and observer). We will use cluster centring within actor to ensure that meta- and observer-perceptions are anchored to the mean validation criterion within each actor (i.e., meta-perceptions will be true-mean centred on the average observer judgment towards an actor, observer attributions will be true-mean centred on the average self-perception provided by actors, and actors' self-perceptions will be mean-centred within actor).

**6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.**

We will exclude any participant who completes the study 'exceptionally fast' (3 SD below the mean). Actor participants who did not describe a TPP/ SPP behaviour, were unclear (e.g., due to containing too little information), or repeatedly answered comprehension questions incorrectly will be removed. We will also exclude observers who incorrectly answer at least one out of two attention check questions.

**7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.**

The sample size is derived it from Lees et al. (2022). Their fourth study – the study which is most similar to the present study - produced a significant moderator of meta-accuracy (our most complex relevant model). Using the "mixedpower" package in R we simulated power for the significant moderator of meta-accuracy model. This data-based estimation used the beta coefficients found in the Study 4 data. Further, actor participant was selected as the random effect to vary in the simulation, with steps between 60-150 as the different sample sizes power should be simulated for. The critical t-value was set to 2, which reflects an alpha level of 5%. The results suggest, that 140 actor participants are sufficient to detect a moderating effect of meta-accuracy of similar size (simulated power = 80%, based on 1000 simulations). We will therefore aim to collect 140 usable TPP behaviours descriptions and 140 usable

SPP behaviour descriptions. We will oversample by 40 participants, anticipating the need for exclusions, resulting in an aim to recruit 180 actor participants.

The sample size for observers is based on the decision to have at least 10 judges rate each behaviour description, and an additional trade-off between a high number of observations and time. Each observer will therefore be asked to rate 8 stories, meaning that 350 observers would be needed so that each of the anticipated 280 descriptions (140 for each behaviour) are rated by around 10 observers. In anticipation of exclusions, we plan to recruit 380 observer participants.

**8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)**

We will investigate potential moderators of observer- and meta-accuracy. Building on the distinctive accuracy model, we will interact the moderator (mean-centred perspective taking, mean-centred empathic-concern, mean-centred personality trait, or punishing type) with distinctive accuracy and normative accuracy.

We predict that some items will be seen as distinctly applicable to TPP or SPP behaviours: observers and punishers will associate TPP more with prosocial motivations and SPP more with self-protection and formidability concerns, and third-party punishers will receive more reputational benefits. We will conduct mixed-effects ANOVAs and perform pairwise comparisons on the estimated marginal means for the interaction between item and punishing type, while adjusting for multiple comparisons using the Tukey method.

We expect punishers to be evaluated more positively in TPP compared to SPP contexts, and for TPP to be perceived as more justified (but positive evaluations of punishers are likely more strongly influenced by the perceived justness of a punishing action than by punishing type). We will fit models predicting mean-centred approval (specifically whether punishing reflected more negatively or more positively on the punisher) as a function of mean-centred justness (specifically as how justified the punishing act was regarded) and punishing type, as well as their interaction, separately for observer and punisher responses.

To assess whether political ideology influences punitive attitudes, we will run six Cumulative Link Mixed Models using the `clmm` function from the `ordinal` package. Hereby, (i) how positively observers believed the punitive action reflected on the punisher and (ii) how justified they believed the punitive action to be will act as the response variable, predicted by either mean-centred (i) left/right political ideology, (ii) RWA, or (iii) SDO, with judge ID as random effect.



## D6 Study Materials for the “Meta-Cognitive and Observer Accuracy Surrounding Punisher Motive Attributions” Study

### Meta-Punishment Stage 1 (Actor)

---

#### Start of Block: info & consent

info *This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, you will be asked to describe two recent instances in which you punished another person for a wrong-doing, before predicting how others might judge you based on those punishing actions.**

The study is estimated to last around 20-25 minutes.

Your submission will be rejected if you do not follow the task instructions when providing behaviour descriptions or if you complete the study exceptionally fast.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. We will collect no identifying information. Completely de-identified datafiles may be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return the survey.

---

#### Page Break

#### notice **Local Data Protection Privacy Notice**

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: ‘Public task’

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return the survey.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)

---

consent Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can message the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
- (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
- (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).
- (5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning the survey.

- ☐ I **consent** to take part in this study (1)
- ☐ I do **NOT consent** to take part in this study (2)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent</strong> to take part in this study*

---

Page Break

---



ID Please enter your Prolific ID

---

Page Break

---

consent 2

**PLEASE READ THE TASK DESCRIPTION CAREFULLY**

This study aims to investigate whether people can accurately predict how others will judge them based on their real-life punishing behaviours. For this, we will conduct a study in two stages. **You are invited to take part in the first stage** of this study.

As part of this study we will ask you to **describe two recent instances in which you punished another person**, before **predicting how others might think of you** based on each description. In the second stage of the study, we will ask another set of participants to read your descriptions, before indicating how they view you based on the punishing behaviour.

Therefore, we would like to use the descriptions you provide as stimuli for participants to rate in future studies. Your behaviour descriptions would only be used for scientific purposes and would be completely anonymised.

***Do you allow us to use your descriptions in future research?***

*Note: If you do not consent, please return the study*

☐ I **consent** (1)

☐ I do **NOT** consent (2)

*Skip To: End of Survey If PLEASE READ THE TASK DESCRIPTION CAREFULLY This study aims to investigate whether people can accu... = I do <strong>NOT </strong>consent*

End of Block: info & consent

---

Start of Block: comprehension

warning Please read the information on the following pages carefully. *If you repeatedly fail the comprehension checks your submission will be rejected.*

-----

Page Break

---

task Later, you will be asked to *describe **two distinct punitive actions you have recently taken***: one in which you punished another when ***you yourself were the victim of an offense***, and another in which you punished a wrong-doer ***after observing the offence as a third party***.

Importantly, we are interested in YOUR punishing behaviour, not the punishing behaviour of other people.

-----

Page Break

---

pun1 In our daily lives we punish people to varying degrees. The wrong-doings that we punish can be serious (e.g. stealing) or mundane (e.g. making an inconsiderate comment), and can be committed by someone familiar or a stranger.

-----

Page Break

---

pun2 Punishment is a behaviour that inflicts a cost on (or removes benefits from) a wrong-doer in response to that persons wrong behaviour.

---

Page Break

---

pun3 Punishment can take many forms, such as directly confronting a wrong-doer (either physically or verbally), sharing negative information about the wrong-doer to others (i.e., gossiping), demanding monetary compensation, discipline (e.g., grounding a child), excluding the wrong-doer from social interactions or avoiding social contact with them.

---

Page Break

---

pun4 You can either punish a wrong-doer after they wronged YOU  
or You can punish a wrong-doer after they wronged SOMEONE ELSE.

You can therefore punish a wrong-doer on behalf of yourself or on behalf of another person.

---

Page Break

---

pun5 An example of punishment *on behalf of yourself* is if you were to stop giving your friend rides because they stole \$50 *from you*.

An example of punishment *on behalf of someone else* is if you were to stop giving your friend rides because they stole \$50 *from another person*.

---

Page Break

---



comp1 Pretend that you have written the following punishing behaviour description: "*I caught my ex vandalising my car, so I demanded they pay me for the damages.*"

Which of the following options matches the description:

- ☐ Punishment on behalf of myself (1)
- ☐ Punishment on behalf of someone else (0)
- ☐ Neither (0)

---

Page Break

*Display This Question:*

*If Pretend that you have written the following punishing behaviour description: "I caught my ex vand... != Punishment on behalf of myself*

incorrect1 Incorrect.

If you had written the description "I caught my ex vandalising my car, so I demanded they pay me for the damages" it would reflect an instance in which you punished a wrong-doing on behalf of yourself, i.e. you were the one affected by the wrong-doing.

---

Page Break

*Display This Question:*

*If Pretend that you have written the following punishing behaviour description: "I caught my ex vand... = Punishment on behalf of myself*

correct1 Correct!

---

Page Break



comp2 Which of the behaviours below is a description of punishment on behalf of someone else:

- ☐ "I was being harassed at the gym so I reported the harasser." (0)
- ☐ "I saw a girl being harassed at the gym so I reported the harasser." (1)
- ☐ Neither description (0)

---

Page Break

*Display This Question:*

*If Which of the behaviours below is a description of punishment on behalf of someone else: != "I saw a girl being harassed at the gym so I reported the harasser."*

incorrect2 Incorrect.

In the scenario "I saw a girl being harassed at the gym so I reported the harasser", the wrong-doer was punished on behalf of someone else (i.e., not you, but another person, was the victim of a wrong-doing).

---

Page Break

*Display This Question:*

*If Which of the behaviours below is a description of punishment on behalf of someone else: = "I saw a girl being harassed at the gym so I reported the harasser."*

correct2 Correct!

---

Page Break



comp3 Which of the following behaviours is a description of punishment on behalf of yourself:

- ☐ I was bored so I went to annoy my sibling (0)
- ☐ A coworker was taking credit for my work so I told other people about it to damage the coworkers reputation (1)
- ☐ Both describe punishment of a wrong-doing on behalf of myself (0)

---

Page Break

*Display This Question:*

*If Which of the following behaviours is a description of punishment on behalf of yourself: != A coworker was taking credit for my work so I told other people about it to damage the coworkers reputation*

incorrect3 Incorrect.

"I was bored so I went to annoy my sibling" does not involve any wrong-doing which can be punished, making the middle option the correct answer.

---

Page Break

*Display This Question:*

*If Which of the following behaviours is a description of punishment on behalf of yourself: = A coworker was taking credit for my work so I told other people about it to damage the coworkers reputation*

correct3 Correct.

The first option ("I was bored so I went to annoy my sibling") does not involve any wrong-doing which could be punished.

---

Page Break



comp4 Which of the following behaviours below is a description in which you punished someone on behalf of someone else:

- ☐ I was bullying someone until someone else came and confronted me. (0)
- ☐ I was being bullied until someone came and confronted them. (0)
- ☐ I was being bullied until I decided to confront the bullies. (0)
- ☐ None of the above (1)

---

Page Break

*Display This Question:*

*If Which of the following behaviours below is a description in which you punished someone on behalf... != None of the above*

incorrect4 Incorrect.

The incorrect options described you as the wrong-doer, someone else punishing the wrong-doer on your behalf, and you punishing on behalf of yourself. A description in which you punish on behalf of someone else would have been "I saw someone being bullied so I went and confronted the bullies."

---

Page Break

*Display This Question:*

*If Which of the following behaviours below is a description in which you punished someone on behalf... = None of the above*

correct4 Correct.

The incorrect options described you as the wrong-doer, someone else punishing the wrong-doer on your behalf, and you punishing on behalf of yourself. A description in which you punish on behalf of someone else would have been "I saw someone being bullied so I went and confronted the bullies."

End of Block: comprehension

---

Start of Block: SPP

SPP

**Please describe in 2-5 sentences, ONE recent instance in which YOU punished someone for engaging in a wrong behaviour that affected YOU.**



After providing your description, you will be asked to provide **self-perceptions** AND **predictions for how others will evaluate you** for this punishing action. Please do not include any identifying information in your writing. However, *please be clear and specific enough so that future participants will understand your description*. **Providing context is encouraged.**

Note that future participants will see the behaviour exactly as you describe it below, without any changes made.

---



---

End of Block: SPP

Start of Block: SPP\_meta



SPPMetaT Imagine *someone else* – the average person – reads your described behaviour (as shown below) and has to rate how they view you based on this. Please indicate how you think **they** would rate your traits based on this described behaviour.

The behaviour: “*\_\${SPP/ChoiceTextEntryValue}\_*”

|                  | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| competent (4)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| cooperative (8)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| dominant (16)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (17)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| moral (18)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| strong (19)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| trustworthy (20) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |



SPPMetaM Imagine *someone else* – the average person – reads your described behaviour (as shown below) and has to rate how much they think your behaviour was motivated by various factors. Please indicate how you think **they** would rate your motivations based on your described behaviour.

The behaviour: “*\_\${SPP/ChoiceTextEntryValue}\_*”

|                                           | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|-------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| appearing<br>cooperative (1)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| asserting<br>authority (40)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| being helpful<br>(41)                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| competitiveness<br>(42)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| defending<br>values (43)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| detering future<br>misconduct (44)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| doing the right<br>thing (45)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| enforcing rules<br>or regulations<br>(46) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| expressing<br>disapproval (47)            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| gaining respect<br>or admiration<br>(48)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| jealousy (49)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| maintaining<br>social order (50)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |

|                                    |                       |                       |                       |                       |                       |                       |                       |
|------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| protecting someone else (52)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| restoring balance or fairness (53) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking justice (54)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking revenge (55)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selfishness (56)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selflessness (57)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| self-protection (62)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| showing strength (58)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| signalling dominance (59)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| spite (60)                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| teaching a lesson (61)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break

reflectionSPPmeta Do you think your punitive action reflected more negatively or more positively on you?

|     | 1.<br>Definitely<br>negatively<br>(1) | 2. (2)                | 3. (3)                | 4. (4)                | 5. (5)                | 6. (6)                | 7.<br>Definitely<br>positively<br>(7) |
|-----|---------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|---------------------------------------|
| (1) | <input type="radio"/>                 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                 |

justifiedSPPmeta Do you think *the average person* would think your punitive action was justified?

|     | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|-----|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: SPP\_meta

Start of Block: SPP\_self



SPPSelfT Based on your described behaviour (as shown below), how would **you** rate yourself on the following traits?

The behaviour: “*\_\${SPP/ChoiceTextEntryValue}\_*”

|                  | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| competent (4)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| cooperative (7)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| dominant (8)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (9)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| moral (16)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| strong (17)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| trustworthy (18) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |



SPPSelfM When engaging in your described behaviour (as shown below), how much were **you** motivated by the following factors?

The behaviour: “*#{SPP/ChoiceTextEntryValue}*”

|                                           | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|-------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| appearing<br>cooperative (1)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| asserting<br>authority (40)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| being helpful<br>(41)                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| competitiveness<br>(42)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| defending<br>values (43)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| detering future<br>misconduct (44)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| doing the right<br>thing (45)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| enforcing rules<br>or regulations<br>(46) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| expressing<br>disapproval (47)            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| gaining respect<br>or admiration<br>(48)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| jealousy (49)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| maintaining<br>social order (50)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |

|                                    |                       |                       |                       |                       |                       |                       |                       |
|------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| protecting someone else (52)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| restoring balance or fairness (53) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking justice (54)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking revenge (55)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selfishness (56)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selflessness (57)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| self-protection (62)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| showing strength (58)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| signalling dominance (59)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| spite (60)                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| teaching a lesson (61)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

---

Page Break

justifiedSPPself Do *you* think your punitive action was justified?

|     | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|-----|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| (4) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: SPP\_self

## Start of Block: TPP

TPP

**Please describe in 2-5 sentences, ONE recent instance in which YOU punished someone for engaging in a wrong behaviour that affected SOMEBODY ELSE.**

After providing your description, you will be asked to provide **self-perceptions** AND **predictions for how others will evaluate you** for this punishing action. Please do not include any identifying information in your writing. However, *please be clear and specific enough so that future participants will understand your description.* **Providing context is encouraged.**

Note that future participants will see the behaviour exactly as you describe it below, without any changes made.

---

---

## End of Block: TPP

## Start of Block: TPP\_meta



TPPMetaT Imagine *someone else* – the average person – reads your described behaviour (as shown below) and has to rate how they view you based on this. Please indicate how you think **they** would rate your traits based on this described behaviour.

The behaviour: “*\${TPP/ChoiceTextEntryValue}*”

|                 | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|-----------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| competent (4)   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| cooperative (7) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| dominant (8)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (9)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| moral (16)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| strong (17)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

trustworthy  
(18)

☐☐☐☐☐☐☐

Page Break



TPPMetaM Imagine *someone else* – the average person – reads your described behaviour (as shown below) and has to rate how much they think your behaviour was motivated by various factors. Please indicate how you think **they** would rate your motivations based on your described behaviour.

The behaviour: “*\$\_{TPP/ChoiceTextEntryValue}*”

|                                           | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|-------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| appearing<br>cooperative (1)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| asserting<br>authority (40)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| being helpful<br>(41)                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| competitiveness<br>(42)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| defending<br>values (43)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| detering future<br>misconduct (44)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| doing the right<br>thing (45)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| enforcing rules<br>or regulations<br>(46) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| expressing<br>disapproval (47)            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| gaining respect<br>or admiration<br>(48)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| jealousy (49)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |



|                                    |                       |                       |                       |                       |                       |                       |                       |
|------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| maintaining social order (50)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| protecting someone else (52)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| restoring balance or fairness (53) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking justice (54)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking revenge (55)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selfishness (56)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selflessness (57)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| self-protection (62)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| showing strength (58)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| signalling dominance (59)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| spite (60)                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| teaching a lesson (61)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break

reflectionTPPmeta Do you think your punitive action reflected more negatively or more positively on you?

|     | 1.<br>Definitely<br>negatively<br>(1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Definitely<br>positively<br>(7) |
|-----|---------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|---------------------------------------|
| (1) | <input type="radio"/>                 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                 |

---

Page Break

justifiedTPPmeta Do you think *the average person* would think your punitive action was justified?

|     | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|-----|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

---

Page Break



attention Do please select completely, the very last button, to show that you are paying attention.

|     | 1. Not at all (0)     | 2 (0)                 | 3 (0)                 | 4 (0)                 | 5 (0)                 | 6 (0)                 | 7. Completely (1)     |
|-----|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: TPP\_meta

---

Start of Block: TPP\_self



unSelfTrait Based on your described behaviour (as shown below), how would **you** rate yourself on the following traits?

The behaviour: “*\${TPP/ChoiceTextEntryValue}*”

|                  | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| competent (4)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| cooperative (7)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| dominant (8)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (9)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| moral (16)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| strong (17)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| trustworthy (18) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break



unSelfMotive When engaging in your described behaviour (as shown below), how much were **you** motivated by the following factors?

The behaviour: “*\${TPP/ChoiceTextEntryValue}*”

|                           | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|---------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| appearing cooperative (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| asserting authority (40)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| being helpful (41)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                                           |                       |                       |                       |                       |                       |                       |                       |
|-------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| competitiveness<br>(42)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| defending<br>values (43)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| detering future<br>misconduct (44)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| doing the right<br>thing (45)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| enforcing rules<br>or regulations<br>(46) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| expressing<br>disapproval (47)            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| gaining respect<br>or admiration<br>(48)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| jealousy (49)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| maintaining<br>social order (50)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| protecting<br>someone else<br>(52)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| restoring<br>balance or<br>fairness (53)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking justice<br>(54)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking revenge<br>(55)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selfishness (56)                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selflessness (57)                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| self-protection<br>(62)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| showing<br>strength (58)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                           |                       |                       |                       |                       |                       |                       |                       |
|---------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| signalling dominance (59) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| spite (60)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| teaching a lesson (61)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break

justifiedTPPself Do *you* think your punitive action was justified?

|     | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|-----|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: TPP\_self

Start of Block: BFIs



BFI How well do the following statements describe your personality?

I see myself as someone who...

|                                        | Disagree strongly (1) | Disagree a little (2) | Neither agree nor disagree (3) | Agree a little (4)    | Agree strongly (5)    |
|----------------------------------------|-----------------------|-----------------------|--------------------------------|-----------------------|-----------------------|
| ...is outgoing, sociable (1)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> |
| ...is reserved (4)                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> |
| ...is generally trusting (5)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> |
| ...tends to find fault with others (6) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> |
| ...does a thorough job (7)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>          | <input type="radio"/> | <input type="radio"/> |

|                                         |                       |                       |                       |                       |                       |
|-----------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| ...tends to be lazy (8)                 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| ...gets nervous easily (9)              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| ...is relaxed, handles stress well (10) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| ...has an active imagination (11)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| ...has few artistic interests (12)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

End of Block: BFIs

Start of Block: Empathic Concern



EC For each item, indicate how well it describes you. Responses range from "Does not describe me well" to "Describes me very well".

|                                                                                          | 1. Does not describe me well (1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5. Describes me very well (5) |
|------------------------------------------------------------------------------------------|----------------------------------|-----------------------|-----------------------|-----------------------|-------------------------------|
| I often have tender, concerned feelings for people less fortunate than me. (1)           | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| Sometimes I don't feel very sorry for other people when they are having problems. (4)    | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| When I see someone being taken advantage of, I feel kind of protective towards them. (5) | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |

|                                                                                                |                       |                       |                       |                       |                       |
|------------------------------------------------------------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Other people's misfortunes do not usually disturb me a great deal. (6)                         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| When I see someone being treated unfairly, I sometimes don't feel very much pity for them. (7) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I am often quite touched by things that I see happen. (8)                                      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| I would describe myself as a pretty soft-hearted person. (9)                                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

#### End of Block: Empathic Concern

#### Start of Block: Perspective Taking



PT For each item, indicate how well it describes you. Responses range from "Does not describe me well" to "Describes me very well".

|                                                                                       | 1. Does not describe me well (1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5. Describes me very well (5) |
|---------------------------------------------------------------------------------------|----------------------------------|-----------------------|-----------------------|-----------------------|-------------------------------|
| I sometimes find it difficult to see things from the "other guy's" point of view. (1) | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| I try to look at everybody's side of a disagreement before I make a decision. (4)     | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| I sometimes try to understand my friends                                              | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |

better by  
imagining how  
things look from  
their  
perspective. (5)

If I'm sure I'm  
right about  
something, I  
don't waste  
much time  
listening to  
other people's  
arguments. (6)

I believe that  
there are two  
sides to every  
question and try  
to look at them  
both. (7)

When I'm upset  
at someone, I  
usually try to  
"put myself in  
his shoes" for a  
while. (8)

Before  
criticizing  
somebody, I try  
to imagine how  
I would feel if I  
were in their  
place. (9)

☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

## End of Block: Perspective Taking

---

### Start of Block: demographics

gender What is your gender?

☐ Male (1)

☐ Female (2)

☐ Other (3)

☐ Prefer not to say (4)





age What is your age in years?

---

End of Block: demographics

---

## Meta-Punishment Stage 2 (Observer)

---

### Start of Block: info & consent

info *This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, you will be asked to read descriptions about recent punishing behaviours prior participants have engaged in. Your task is to indicate how you perceive the authors based on their described behaviour.**

The study is estimated to last around 15-20 minutes.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. We will collect no identifying information. Completely de-identified datafiles may be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return the survey.

---

Page Break

---

### notice **Local Data Protection Privacy Notice**

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: 'Public task'

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return the survey.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)

---

consent Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can message the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
- (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
- (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).
- (5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning the survey.

☐ I **consent** to take part in this study (1)

☐ I do **NOT consent** to take part in this study (2)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent</strong> to take part in this study*

Page Break



ID Please enter your Prolific ID

Page Break

task

**PLEASE READ THE TASK DESCRIPTION CAREFULLY**

This study aims to investigate whether people can accurately predict how others will judge them based on their real-life behaviours. For this, we are conducting a study in two stages. **You are invited to take part in the second stage** of this study.

In the first stage we asked participants to describe recent punishing behaviours they engaged in, and to then predict how others will view them based on the description. Now, in the second stage of the

study, we will ask you to **read eight descriptions provided by eight prior participants**, and to **indicate how you perceive those participants based on their behaviour**.

---

Page Break

---

context1 In our daily lives we punish people to varying degrees. The wrong-doings that we punish can be serious (e.g. stealing) or mundane (e.g. making an inconsiderate comment), and can be committed by someone familiar or a stranger.

---

Page Break

---

context2 Punishment is a behaviour that inflicts a cost on (or removes benefits from) a wrong-doer in response to that persons wrong behaviour.

---

Page Break

---

context3 Punishment can take many forms, such as directly confronting a wrong-doer (either physically or verbally), sharing negative information about the wrong-doer to others (i.e., gossiping), demanding monetary compensation, discipline (e.g., grounding a child), excluding the wrong-doer from social interactions or avoiding social contact with them.

End of Block: info & consent

---

Start of Block: TPP subset



TPPaction TPP behaviours  
*[actor responses]*

End of Block: TPP subset

---

Start of Block: TPP rate



TPPtrait Below is a description of a punishing behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, **indicate how you perceive the author**.

“\${Im://Field/1}”

|                     | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|---------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| competent<br>(4)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| cooperative<br>(7)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| dominant<br>(8)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| likeable (9)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| moral (16)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| strong (17)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| trustworthy<br>(18) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |



TPPmotive Below is the same description of a punishing behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, **indicate what you believe the author's punishing motives were.**

“\${Im://Field/1}”

|                              | 1. Not at<br>all (1)  | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7.<br>Completely<br>(7) |
|------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| appearing<br>cooperative (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| asserting<br>authority (18)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| being helpful<br>(19)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |
| competitiveness<br>(20)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>   |

|                                     |                       |                       |                       |                       |                       |                       |                       |
|-------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| defending values (21)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| detering future misconduct (22)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| doing the right thing (23)          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| enforcing rules or regulations (24) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| expressing disapproval (25)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| gaining respect or admiration (26)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| jealousy (27)                       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| maintaining social order (28)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| protecting someone else (29)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| restoring balance or fairness (30)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking justice (31)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking revenge (32)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selfishness (33)                    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selflessness (34)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| self-protection (35)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| showing strength (36)               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| signalling dominance (37)           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                        |                       |                       |                       |                       |                       |                       |                       |
|------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| spite (38)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| teaching a lesson (39) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break

TPPjustified Do you think the punitive action was justified?

|     |                       |                       |                       |                       |                       |                       |                       |
|-----|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|     | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
| (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break

TPPreflection Do you think the punitive action reflected more negatively or more positively on the author?

|     |                              |                       |                       |                       |                       |                       |                              |
|-----|------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------------|
|     | 1. Definitely negatively (1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Definitely positively (7) |
| (1) | <input type="radio"/>        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>        |

End of Block: TPP rate

Start of Block: SPP subset



SPPaction SPP behaviours  
[actor responses]

End of Block: SPP subset

Start of Block: SPP rate



SPPtrait Below is a description of a punishing behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, **indicate how you perceive the author**.

“ $\{lm://Field/1\}$ ”

|                  | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| competent (4)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| cooperative (7)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| dominant (8)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| likeable (9)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| moral (16)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| strong (17)      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| trustworthy (18) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |



SPPmotive Below is the same description of a punishing behaviour a previous participant engaged in. Please read the description of their recent action and, based on this, **indicate what you believe the author's punishing motives were**.

“ $\{lm://Field/1\}$ ”

|                           | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
|---------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| appearing cooperative (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| asserting authority (18)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |



|                                           |                       |                       |                       |                       |                       |                       |                       |
|-------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| being helpful<br>(19)                     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| competitiveness<br>(20)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| defending<br>values (21)                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| detering future<br>misconduct<br>(22)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| doing the right<br>thing (23)             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| enforcing rules<br>or regulations<br>(24) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| expressing<br>disapproval<br>(25)         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| gaining respect<br>or admiration<br>(26)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| jealousy (27)                             | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| maintaining<br>social order<br>(28)       | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| protecting<br>someone else<br>(29)        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| restoring<br>balance or<br>fairness (30)  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking justice<br>(31)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| seeking revenge<br>(32)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selfishness (33)                          | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| selflessness<br>(34)                      | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| self-protection<br>(35)                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

|                           |                       |                       |                       |                       |                       |                       |                       |
|---------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| showing strength (36)     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| signalling dominance (37) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| spite (38)                | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| teaching a lesson (39)    | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break

SPPjustified Do you think the punitive action was justified?

|     |                       |                       |                       |                       |                       |                       |                       |
|-----|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|     | 1. Not at all (1)     | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Completely (7)     |
| (1) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page Break

SPPreflection Do you think the punitive action reflected more negatively or more positively on the author?

|     |                              |                       |                       |                       |                       |                       |                              |
|-----|------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------------|
|     | 1. Definitely negatively (1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5 (5)                 | 6 (6)                 | 7. Definitely positively (7) |
| (1) | <input type="radio"/>        | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>        |

End of Block: SPP rate

Start of Block: attention1



attention1 Do please select definitely positively, the very last button, to show that you are paying attention.

|     | 1.<br>Definitely<br>negatively<br>(0) | 2 (0)                 | 3 (0)                 | 4 (0)                 | 5 (0)                 | 6 (0)                 | 7.<br>Definitely<br>positively<br>(1) |
|-----|---------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|---------------------------------------|
| (1) | <input type="radio"/>                 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>                 |

End of Block: attention1

Start of Block: BFIs



BFI How well do the following statements describe your personality?

I see myself as someone who...

|                                               | Disagree<br>strongly (1) | Disagree a little<br>(2) | Neither agree<br>nor disagree<br>(3) | Agree a little<br>(4) | Agree strongly<br>(5) |
|-----------------------------------------------|--------------------------|--------------------------|--------------------------------------|-----------------------|-----------------------|
| ...is outgoing,<br>sociable (1)               | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| ...is reserved (4)                            | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| ...is generally<br>trusting (5)               | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| ...tends to find<br>fault with<br>others (6)  | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| ...does a<br>thorough job<br>(7)              | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| ...tends to be<br>lazy (8)                    | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| ...gets nervous<br>easily (9)                 | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| ...is relaxed,<br>handles stress<br>well (10) | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |
| ...has an active<br>imagination<br>(11)       | <input type="radio"/>    | <input type="radio"/>    | <input type="radio"/>                | <input type="radio"/> | <input type="radio"/> |

...has few  
artistic interests  
(12)

☐ ☐ ☐ ☐ ☐

End of Block: BFIs

Start of Block: Empathic Concern



EC For each item, indicate how well it describes you. Responses range from "Does not describe me well" to "Describes me very well".

|                                                                                                                  | 1. Does not<br>describe me<br>well (1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5. Describes me<br>very well (5) |
|------------------------------------------------------------------------------------------------------------------|----------------------------------------|-----------------------|-----------------------|-----------------------|----------------------------------|
| I often have<br>tender,<br>concerned<br>feelings for<br>people less<br>fortunate than<br>me. (1)                 | <input type="radio"/>                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>            |
| Sometimes I<br>don't feel very<br>sorry for other<br>people when<br>they are having<br>problems. (4)             | <input type="radio"/>                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>            |
| When I see<br>someone being<br>taken<br>advantage of, I<br>feel kind of<br>protective<br>towards them.<br>(5)    | <input type="radio"/>                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>            |
| Other people's<br>misfortunes do<br>not usually<br>disturb me a<br>great deal. (6)                               | <input type="radio"/>                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>            |
| When I see<br>someone being<br>treated unfairly,<br>I sometimes<br>don't feel very<br>much pity for<br>them. (7) | <input type="radio"/>                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>            |

I am often quite touched by things that I see happen. (8)

☐ ☐ ☐ ☐ ☐

I would describe myself as a pretty soft-hearted person. (9)

☐ ☐ ☐ ☐ ☐

#### End of Block: Empathic Concern

#### Start of Block: Perspective Taking



PT For each item, indicate how well it describes you. Responses range from "Does not describe me well" to "Describes me very well".

|                                                                                                           | 1. Does not describe me well (1) | 2 (2)                 | 3 (3)                 | 4 (4)                 | 5. Describes me very well (5) |
|-----------------------------------------------------------------------------------------------------------|----------------------------------|-----------------------|-----------------------|-----------------------|-------------------------------|
| I sometimes find it difficult to see things from the "other guy's" point of view. (1)                     | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| I try to look at everybody's side of a disagreement before I make a decision. (4)                         | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| I sometimes try to understand my friends better by imagining how things look from their perspective. (5)  | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |
| If I'm sure I'm right about something, I don't waste much time listening to other people's arguments. (6) | <input type="radio"/>            | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>         |

I believe that there are two sides to every question and try to look at them both. (7)

☐ ☐ ☐ ☐ ☐

When I'm upset at someone, I usually try to "put myself in his shoes" for a while. (8)

☐ ☐ ☐ ☐ ☐

Before criticizing somebody, I try to imagine how I would feel if I were in their place. (9)

☐ ☐ ☐ ☐ ☐

#### End of Block: Perspective Taking

#### Start of Block: Right Wing Authoritarianism



RWA Please indicate the degree to which you agree or disagree with each statement. Responses range from "very strongly disagree" to "very strongly agree".

|                                                                                              | very<br>strongl<br>y<br>disagre<br>e (1) | strongl<br>y<br>disagre<br>e (2) | somewh<br>at<br>disagree<br>(3) | slightly<br>disagre<br>e (4) | neither<br>agree<br>nor<br>disagre<br>e (5) | slightl<br>y<br>agree<br>(6) | somewh<br>at agree<br>(7) | strongl<br>y<br>agree<br>(8) | very<br>strongl<br>y<br>agree<br>(9) |
|----------------------------------------------------------------------------------------------|------------------------------------------|----------------------------------|---------------------------------|------------------------------|---------------------------------------------|------------------------------|---------------------------|------------------------------|--------------------------------------|
| It's great that many young people today are prepared to defy authority. (1)                  | <input type="radio"/>                    | <input type="radio"/>            | <input type="radio"/>           | <input type="radio"/>        | <input type="radio"/>                       | <input type="radio"/>        | <input type="radio"/>     | <input type="radio"/>        | <input type="radio"/>                |
| What our country needs most is discipline, with everyone following our leaders in unity. (4) | <input type="radio"/>                    | <input type="radio"/>            | <input type="radio"/>           | <input type="radio"/>        | <input type="radio"/>                       | <input type="radio"/>        | <input type="radio"/>     | <input type="radio"/>        | <input type="radio"/>                |
| God's laws about                                                                             | <input type="radio"/>                    | <input type="radio"/>            | <input type="radio"/>           | <input type="radio"/>        | <input type="radio"/>                       | <input type="radio"/>        | <input type="radio"/>     | <input type="radio"/>        | <input type="radio"/>                |

abortion,  
pornography,  
and marriage  
must be  
strictly  
followed  
before it is  
too late. (5)

There is  
nothing  
wrong with  
premarital  
sexual  
intercourse.  
(6)

Our society  
does NOT  
need tougher  
government  
and stricter  
laws. (7)

The facts on  
crime and  
the recent  
public  
disorders  
show we  
have to crack  
down harder  
on  
troublemake  
rs, if we are  
going to  
preserve law  
and order.  
(8)

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

## End of Block: Right Wing Authoritarianism

## Start of Block: Social Dominance Orientation



SDO Show how much you favour or oppose each idea below by selecting a number from 1 to 7 on the scale below. You can work quickly; your first feeling is generally best.

|                     | 1. Strongly<br>Oppose<br>(1) | 2.<br>Somewhat<br>Oppose (2) | 3. Slightly<br>Oppose<br>(3) | 4. Neutral<br>(4)     | 5. Slightly<br>Favour (5) | 6.<br>Somewhat<br>Favour (6) | 7. Strongly<br>Favour (7) |
|---------------------|------------------------------|------------------------------|------------------------------|-----------------------|---------------------------|------------------------------|---------------------------|
| An ideal<br>society | <input type="radio"/>        | <input type="radio"/>        | <input type="radio"/>        | <input type="radio"/> | <input type="radio"/>     | <input type="radio"/>        | <input type="radio"/>     |

requires  
some  
groups to  
be on top  
and others  
to be on  
the  
bottom.  
(1)

Some  
groups of  
people are  
simply  
inferior to  
other  
groups.  
(4)

No one  
group  
should  
dominate  
in society.  
(5)

Groups at  
the bottom  
are just as  
deserving  
as groups  
at the top.  
(6)

Group  
equality  
should not  
be our  
primary  
goal. (7)

It is unjust  
to try to  
make  
groups  
equal. (8)

We should  
do what  
we can to  
equalize  
conditions  
for  
different  
groups. (9)

We should  
work to  
give all

|                       |                       |                       |                       |                       |                       |                       |                       |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

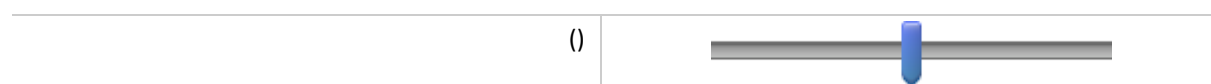


## End of Block: Social Dominance Orientation



Left                  Centre/Moderate                  Right

0                                                  100



Start of Block: attention2



- ☐ punitive actions (1)
- ☐ consumer behaviours (0)
- ☐ dietary preferences (0)
- ☐ scary stories (0)

## Start of Block: demographics

gender What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



age What is your age in years?

---

End of Block: demographics

---

## D7 Agreement on the Relative Importance of Motives

One aim of Chapter 5 was to assess which motives are differentially associated with TPP versus SPP. While the Results section presents three models predicting mean-centred (observer-, meta-, and self-perception) ratings as a function of item, punishing type and their interaction, the following presents relative position plots (Figures D7.1 & D7.2). These plots illustrate the agreement between observer and punisher ratings regarding the relative importance or applicability of motives in second-party (Figure D7.1) and third-party punishment (Figure D7.2) situations. These plots were created to visualise how the relative positions of motives match between observer-, meta-, and self-perceptions.

**Figure D7.1**

*Agreement on the Relative Importance of Motives in SPP: Observer, Meta-, and Self-Perceptions*

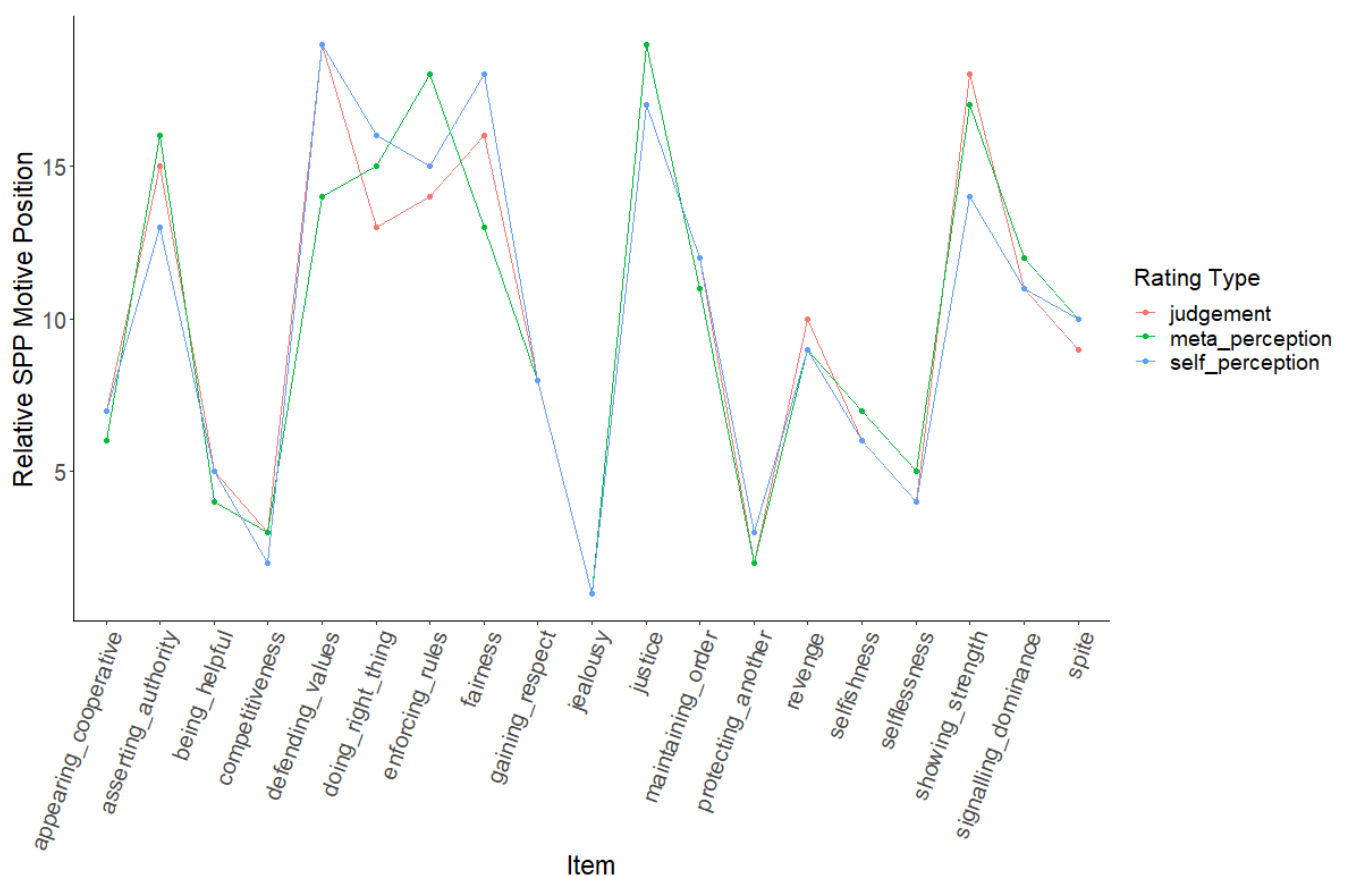
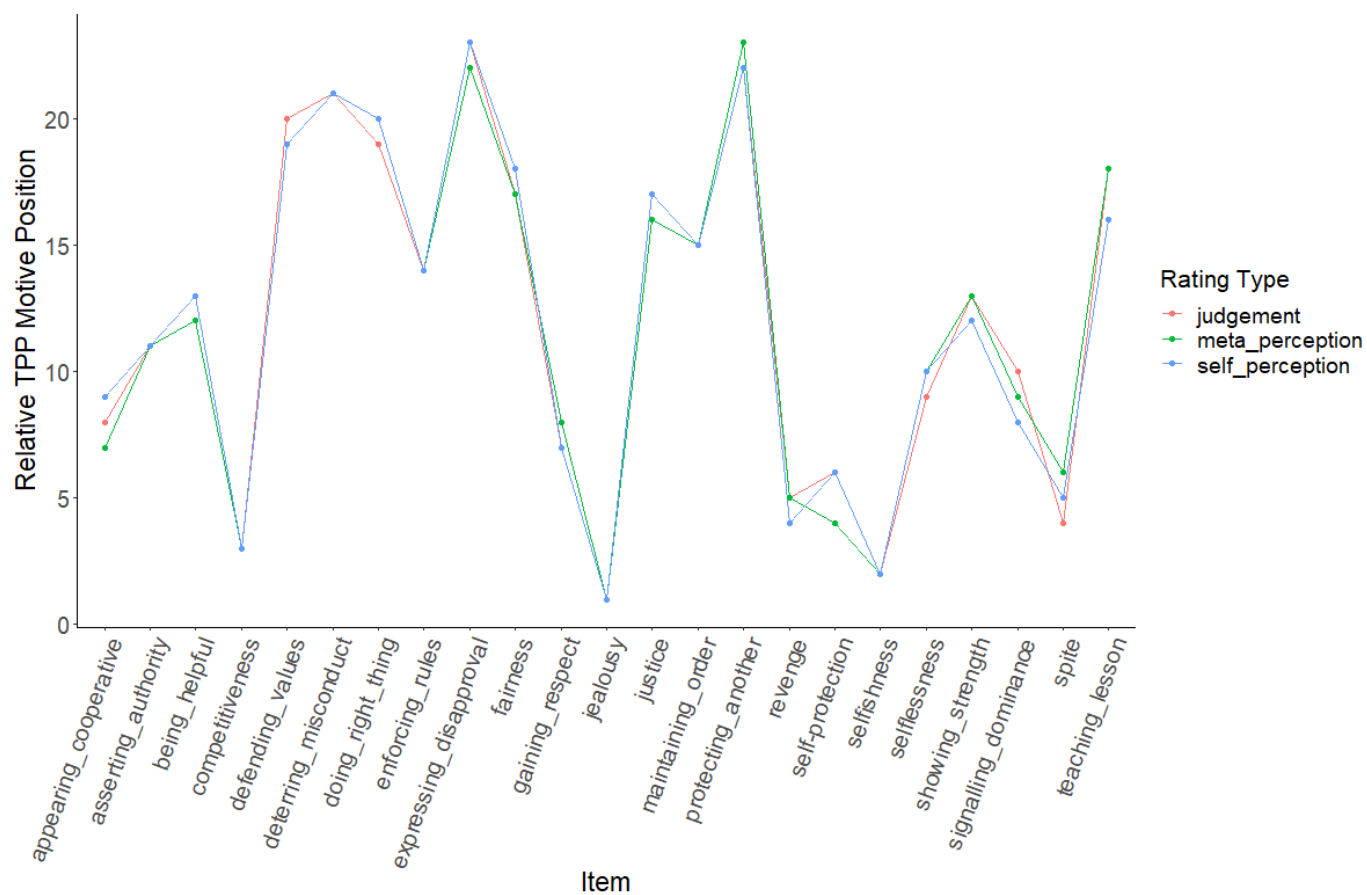


**Figure D7.2**

*Agreement on the Relative Importance of Motives in TPP: Observer, Meta-, and Self-Perceptions*



## D8 Correlation Matrices of Traits and Motives for SPP and TPP Behaviours

**Table D8.1**

*Correlation Matrix of Traits and Motives for SPP Behaviours*

|                            | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9    | 10    | 11   | 12    | 13    | 14    | 15    | 16    | 17    | 18    | 19    | 20    | 21    | 22    | 23    | 24    | 25    | 26    | 27    | 28   | 29   | 30 |
|----------------------------|-------|-------|-------|-------|-------|-------|-------|-------|------|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|------|----|
| 1. competent               | --    |       |       |       |       |       |       |       |      |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 2. cooperative             | 0.53  | --    |       |       |       |       |       |       |      |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 3. dominant                | 0.44  | 0.33  | --    |       |       |       |       |       |      |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 4. likeable                | 0.69  | 0.67  | 0.38  | --    |       |       |       |       |      |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 5. moral                   | 0.75  | 0.61  | 0.31  | 0.74  | --    |       |       |       |      |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 6. strong                  | 0.7   | 0.51  | 0.62  | 0.65  | 0.59  | --    |       |       |      |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 7. trustworthy             | 0.78  | 0.67  | 0.41  | 0.78  | 0.8   | 0.7   | --    |       |      |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 8. appearing cooperative   | 0.39  | 0.76  | 0.31  | 0.53  | 0.48  | 0.38  | 0.5   | --    |      |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 9. asserting authority     | 0.36  | 0.27  | 0.65  | 0.29  | 0.28  | 0.48  | 0.3   | 0.3   | --   |       |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 10. being helpful          | 0.4   | 0.61  | 0.26  | 0.58  | 0.45  | 0.35  | 0.47  | 0.67  | 0.22 | --    |      |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 11. competitiveness        | 0.07  | 0.27  | 0.27  | 0.08  | 0.03  | 0.18  | 0.05  | 0.37  | 0.3  | 0.32  | --   |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 12. defending values       | 0.61  | 0.55  | 0.4   | 0.68  | 0.66  | 0.54  | 0.69  | 0.53  | 0.32 | 0.49  | 0.2  | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 13. deterring misconduct   | 0.62  | 0.45  | 0.43  | 0.55  | 0.6   | 0.57  | 0.61  | 0.39  | 0.42 | 0.35  | 0.01 | 0.53  | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 14. doing the right thing  | 0.65  | 0.59  | 0.34  | 0.67  | 0.77  | 0.57  | 0.72  | 0.53  | 0.25 | 0.54  | 0.06 | 0.67  | 0.6   | --    |       |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 15. enforcing rules        | 0.57  | 0.48  | 0.34  | 0.41  | 0.53  | 0.43  | 0.49  | 0.47  | 0.48 | 0.38  | 0.2  | 0.48  | 0.57  | 0.6   | --    |       |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 16. expressing disapproval | 0.26  | 0.08  | 0.26  | 0.2   | 0.28  | 0.27  | 0.25  | 0     | 0.24 | 0.15  | 0.09 | 0.26  | 0.37  | 0.22  | 0.18  | --    |       |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 17. gaining respect        | 0.37  | 0.52  | 0.44  | 0.55  | 0.41  | 0.44  | 0.5   | 0.61  | 0.44 | 0.56  | 0.24 | 0.55  | 0.45  | 0.48  | 0.38  | 0.08  | --    |       |       |       |       |       |       |       |       |       |       |      |      |    |
| 18. jealousy               | -0.16 | 0.09  | 0.14  | -0.02 | -0.08 | 0.02  | -0.05 | 0.13  | 0.2  | 0.13  | 0.48 | 0.1   | -0.12 | 0     | 0.03  | 0.1   | 0.11  | --    |       |       |       |       |       |       |       |       |       |      |      |    |
| 19. maintaining order      | 0.54  | 0.49  | 0.39  | 0.47  | 0.51  | 0.47  | 0.53  | 0.5   | 0.44 | 0.42  | 0.27 | 0.61  | 0.5   | 0.57  | 0.62  | 0.16  | 0.45  | 0.1   | --    |       |       |       |       |       |       |       |       |      |      |    |
| 20. protecting someone     | 0.16  | 0.35  | 0.18  | 0.28  | 0.22  | 0.14  | 0.23  | 0.45  | 0.08 | 0.59  | 0.18 | 0.36  | 0.29  | 0.36  | 0.23  | 0.08  | 0.36  | 0.1   | 0.25  | --    |       |       |       |       |       |       |       |      |      |    |
| 21. restoring fairness     | 0.55  | 0.47  | 0.38  | 0.49  | 0.63  | 0.53  | 0.52  | 0.48  | 0.29 | 0.39  | 0.14 | 0.63  | 0.52  | 0.61  | 0.5   | 0.17  | 0.43  | -0.01 | 0.58  | 0.26  | --    |       |       |       |       |       |       |      |      |    |
| 22. seeking justice        | 0.4   | 0.33  | 0.28  | 0.29  | 0.39  | 0.4   | 0.4   | 0.27  | 0.22 | 0.26  | 0.24 | 0.45  | 0.28  | 0.42  | 0.37  | 0.18  | 0.23  | 0.08  | 0.41  | 0.14  | 0.42  | --    |       |       |       |       |       |      |      |    |
| 23. seeking revenge        | -0.31 | -0.31 | -0.05 | -0.34 | -0.42 | -0.21 | -0.35 | -0.3  | 0.02 | -0.26 | 0.12 | -0.33 | -0.34 | -0.43 | -0.32 | -0.06 | -0.21 | 0.24  | -0.25 | -0.16 | -0.35 | 0.06  | --    |       |       |       |       |      |      |    |
| 24. selfishness            | -0.27 | -0.23 | 0.06  | -0.31 | -0.4  | -0.21 | -0.38 | -0.17 | 0.06 | -0.09 | 0.24 | -0.14 | -0.3  | -0.4  | -0.26 | -0.1  | -0.09 | 0.25  | -0.22 | 0.02  | -0.19 | -0.04 | 0.44  | --    |       |       |       |      |      |    |
| 25. selflessness           | 0.35  | 0.56  | 0.4   | 0.49  | 0.39  | 0.42  | 0.43  | 0.69  | 0.29 | 0.6   | 0.37 | 0.49  | 0.33  | 0.46  | 0.35  | 0.12  | 0.56  | 0.21  | 0.47  | 0.45  | 0.4   | 0.25  | -0.16 | -0.08 | --    |       |       |      |      |    |
| 26. self-protection        | 0.42  | 0.33  | 0.32  | 0.41  | 0.36  | 0.52  | 0.45  | 0.31  | 0.13 | 0.25  | 0.17 | 0.43  | 0.32  | 0.35  | 0.13  | 0.2   | 0.3   | 0.13  | 0.32  | 0.18  | 0.42  | 0.22  | -0.13 | -0.03 | 0.25  | --    |       |      |      |    |
| 27. showing strength       | 0.59  | 0.41  | 0.65  | 0.57  | 0.49  | 0.76  | 0.61  | 0.37  | 0.55 | 0.37  | 0.19 | 0.54  | 0.57  | 0.54  | 0.45  | 0.31  | 0.48  | 0.12  | 0.45  | 0.23  | 0.51  | 0.41  | -0.13 | -0.16 | 0.38  | 0.47  | --    |      |      |    |
| 28. signalling dominance   | 0.24  | 0.27  | 0.65  | 0.26  | 0.19  | 0.39  | 0.26  | 0.28  | 0.74 | 0.23  | 0.39 | 0.31  | 0.39  | 0.21  | 0.36  | 0.23  | 0.48  | 0.19  | 0.36  | 0.12  | 0.23  | 0.3   | 0.08  | 0.12  | 0.26  | 0.15  | 0.54  | --   |      |    |
| 29. spite                  | -0.3  | -0.3  | 0.03  | -0.31 | -0.37 | -0.15 | -0.33 | -0.24 | 0.06 | -0.22 | 0.17 | -0.26 | -0.27 | -0.4  | -0.3  | 0.08  | -0.12 | 0.22  | -0.29 | -0.11 | -0.31 | -0.03 | 0.74  | 0.52  | -0.14 | -0.07 | -0.06 | 0.18 | --   |    |
| 30. teaching a lesson      | 0.36  | 0.24  | 0.32  | 0.26  | 0.18  | 0.34  | 0.27  | 0.24  | 0.45 | 0.18  | 0.04 | 0.19  | 0.42  | 0.25  | 0.33  | 0.28  | 0.28  | -0.08 | 0.26  | 0.12  | 0.14  | 0.19  | 0.09  | -0.08 | 0.23  | 0.01  | 0.38  | 0.35 | 0.11 | -- |

*Note.* This table displays the Pearson correlation coefficients between traits (items 1–7) and motives (items 8–30) based on participants' meta-perceptions of how their SPP behaviours might be observed by others.

**Table D8.2***Correlation Matrix of Traits and Motives for TPP Behaviours*

|                            | 1     | 2     | 3    | 4     | 5     | 6     | 7     | 8    | 9     | 10    | 11    | 12    | 13    | 14    | 15    | 16   | 17    | 18    | 19    | 20    | 21    | 22   | 23    | 24    | 25    | 26   | 27   | 28   | 29   | 30 |
|----------------------------|-------|-------|------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|------|-------|-------|-------|------|------|------|------|----|
| 1. competent               | --    |       |      |       |       |       |       |      |       |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 2. cooperative             | 0.63  | --    |      |       |       |       |       |      |       |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 3. dominant                | 0.35  | 0.3   | --   |       |       |       |       |      |       |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 4. likeable                | 0.62  | 0.69  | 0.39 | --    |       |       |       |      |       |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 5. moral                   | 0.73  | 0.6   | 0.19 | 0.69  | --    |       |       |      |       |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 6. strong                  | 0.67  | 0.55  | 0.56 | 0.62  | 0.56  | --    |       |      |       |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 7. trustworthy             | 0.76  | 0.64  | 0.33 | 0.75  | 0.78  | 0.68  | --    |      |       |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 8. appearing cooperative   | 0.4   | 0.52  | 0.35 | 0.43  | 0.22  | 0.31  | 0.39  | --   |       |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 9. asserting authority     | 0.3   | 0.31  | 0.64 | 0.27  | 0.18  | 0.42  | 0.3   | 0.37 | --    |       |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 10. being helpful          | 0.47  | 0.62  | 0.22 | 0.6   | 0.56  | 0.5   | 0.54  | 0.52 | 0.3   | --    |       |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 11. competitiveness        | 0.05  | 0.15  | 0.34 | 0.12  | -0.03 | 0.22  | 0.12  | 0.35 | 0.37  | 0.16  | --    |       |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 12. defending values       | 0.48  | 0.5   | 0.19 | 0.56  | 0.57  | 0.44  | 0.52  | 0.31 | 0.25  | 0.52  | 0.15  | --    |       |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 13. deterring misconduct   | 0.25  | 0.35  | 0.3  | 0.43  | 0.3   | 0.26  | 0.35  | 0.28 | 0.38  | 0.42  | 0.08  | 0.34  | --    |       |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 14. doing the right thing  | 0.48  | 0.55  | 0.14 | 0.56  | 0.7   | 0.43  | 0.58  | 0.31 | 0.22  | 0.64  | -0.01 | 0.63  | 0.31  | --    |       |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 15. enforcing rules        | 0.37  | 0.43  | 0.34 | 0.37  | 0.37  | 0.34  | 0.4   | 0.42 | 0.4   | 0.43  | 0.15  | 0.26  | 0.53  | 0.35  | --    |      |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 16. expressing disapproval | 0.29  | 0.13  | 0.19 | 0.16  | 0.17  | 0.31  | 0.25  | 0.1  | 0.28  | 0.2   | 0.25  | 0.3   | 0.08  | 0.13  | 0.05  | --   |       |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 17. gaining respect        | 0.3   | 0.44  | 0.43 | 0.46  | 0.29  | 0.33  | 0.41  | 0.53 | 0.53  | 0.5   | 0.44  | 0.38  | 0.27  | 0.35  | 0.33  | 0.2  | --    |       |       |       |       |      |       |       |       |      |      |      |      |    |
| 18. jealousy               | -0.04 | 0.06  | 0.14 | 0.03  | -0.08 | 0.02  | -0.01 | 0.26 | 0.17  | 0.01  | 0.54  | -0.04 | -0.07 | -0.02 | 0.1   | 0.07 | 0.23  | --    |       |       |       |      |       |       |       |      |      |      |      |    |
| 19. maintaining order      | 0.33  | 0.46  | 0.4  | 0.47  | 0.39  | 0.42  | 0.42  | 0.38 | 0.51  | 0.46  | 0.25  | 0.44  | 0.5   | 0.42  | 0.62  | 0.19 | 0.56  | 0.11  | --    |       |       |      |       |       |       |      |      |      |      |    |
| 20. protecting someone     | 0.22  | 0.34  | 0.21 | 0.4   | 0.37  | 0.34  | 0.27  | 0.21 | 0.2   | 0.46  | 0.16  | 0.38  | 0.34  | 0.42  | 0.15  | 0.16 | 0.35  | 0.05  | 0.31  | --    |       |      |       |       |       |      |      |      |      |    |
| 21. restoring fairness     | 0.54  | 0.51  | 0.38 | 0.51  | 0.47  | 0.57  | 0.54  | 0.45 | 0.45  | 0.56  | 0.15  | 0.48  | 0.44  | 0.5   | 0.47  | 0.22 | 0.41  | -0.03 | 0.5   | 0.28  | --    |      |       |       |       |      |      |      |      |    |
| 22. seeking justice        | 0.41  | 0.27  | 0.28 | 0.35  | 0.29  | 0.43  | 0.37  | 0.31 | 0.31  | 0.42  | 0.18  | 0.45  | 0.33  | 0.38  | 0.3   | 0.31 | 0.17  | 0.14  | 0.3   | 0.23  | 0.47  | --   |       |       |       |      |      |      |      |    |
| 23. seeking revenge        | -0.06 | -0.1  | 0.25 | -0.05 | -0.25 | 0.08  | -0.13 | 0.13 | 0.16  | -0.08 | 0.36  | 0.01  | 0.07  | -0.2  | -0.03 | 0.16 | 0.06  | 0.27  | 0.03  | 0.06  | 0.09  | 0.27 | --    |       |       |      |      |      |      |    |
| 24. selfishness            | -0.04 | -0.06 | 0.03 | -0.08 | -0.18 | -0.03 | -0.13 | 0.18 | -0.11 | -0.01 | 0.22  | -0.02 | -0.04 | -0.12 | -0.1  | 0    | 0.05  | 0.34  | -0.05 | -0.07 | -0.07 | 0.12 | 0.31  | --    |       |      |      |      |      |    |
| 25. selflessness           | 0.39  | 0.5   | 0.34 | 0.57  | 0.47  | 0.44  | 0.5   | 0.44 | 0.31  | 0.62  | 0.23  | 0.48  | 0.28  | 0.46  | 0.2   | 0.15 | 0.55  | 0.07  | 0.38  | 0.43  | 0.42  | 0.28 | -0.01 | 0.09  | --    |      |      |      |      |    |
| 26. self-protection        | 0.24  | 0.22  | 0.21 | 0.29  | 0.14  | 0.28  | 0.26  | 0.37 | 0.17  | 0.17  | 0.26  | 0.21  | 0.07  | 0.15  | 0.18  | 0.03 | 0.19  | 0.28  | 0.07  | 0.09  | 0.17  | 0.19 | 0.17  | 0.28  | 0.24  | --   |      |      |      |    |
| 27. showing strength       | 0.35  | 0.32  | 0.5  | 0.42  | 0.29  | 0.64  | 0.41  | 0.35 | 0.62  | 0.45  | 0.31  | 0.36  | 0.33  | 0.34  | 0.29  | 0.32 | 0.39  | 0.07  | 0.39  | 0.3   | 0.49  | 0.51 | 0.26  | -0.04 | 0.42  | 0.31 | --   |      |      |    |
| 28. signalling dominance   | 0.19  | 0.2   | 0.68 | 0.24  | 0.03  | 0.38  | 0.18  | 0.29 | 0.68  | 0.18  | 0.44  | 0.14  | 0.27  | 0.02  | 0.29  | 0.27 | 0.49  | 0.25  | 0.4   | 0.18  | 0.28  | 0.25 | 0.3   | 0.08  | 0.23  | 0.16 | 0.51 | --   |      |    |
| 29. spite                  | -0.15 | -0.2  | 0.22 | -0.08 | -0.27 | 0.03  | -0.11 | 0.02 | 0.11  | -0.2  | 0.37  | -0.14 | -0.04 | -0.27 | -0.12 | 0.09 | -0.02 | 0.37  | -0.05 | 0     | -0.06 | 0.11 | 0.66  | 0.3   | -0.12 | 0.18 | 0.22 | 0.28 | --   |    |
| 30. teaching a lesson      | 0.33  | 0.26  | 0.44 | 0.31  | 0.18  | 0.36  | 0.42  | 0.35 | 0.43  | 0.27  | 0.18  | 0.17  | 0.43  | 0.15  | 0.35  | 0.23 | 0.3   | -0.04 | 0.36  | 0.13  | 0.46  | 0.34 | 0.2   | 0.03  | 0.2   | 0.11 | 0.35 | 0.39 | 0.21 | -- |

*Note.* This table displays the Pearson correlation coefficients between traits (items 1–7) and motives (items 8–30) based on participants' meta-perceptions of how their TPP behaviours might be observed by others.

## D9 A Sample of Free Text Responses

This section presents a selection of participant responses to illustrate the variety of second- and third-party punishing behaviours reported. The responses reflect a wide range of punitive actions, across different contexts and relationships, with descriptions also varying in level of detail and focus (e.g., explaining situations, justifying behaviours, or highlighting various motivations).

### *Second-Party Punishing Behaviours*

Participants described a variety of second-party punishing actions, including child discipline (*“My 2-year old son threw one of his toys at me because I told him it was time for a bath. I took his toy away and told him it was not okay to throw his toys, especially at people or animals”*), withdrawal of benefits (*“This person took advantage of me while I gave them rides to their work. They promised to compensate me for gas spent and they did not end up paying it to me. I texted them telling them that I would not give them another ride to work until they payed me for the rides I had already gave them”*), physical retaliation (*“One recent instance that I can think of in which I punished somebody for engaging in behavior that affected myself was when my little brother smacked me in the head, so I smacked him right back as punishment. I don't think others would love the punitive action, as it was purely revenge-based. I don't love what I did either, but it was almost an instinct to hit back”*), severing relationships (*“I found out that one of my friends was talking badly about me behind my back so I told her I did not want to be friends any longer and cut all ties with her. She was being a mean person”*), material compensation (*“I gave someone my phone and they dropped it as a joke. The phone cracked and I made them reimburse me the repair costs”*), tattling to parents (*“Teenagers were damaging mailboxes with a bat as they drove by on a four wheeler. I was looking out my window and saw it happen. I knew where they lived, so I went to the home and told the parents about this incident”*), and calling the police (*“My neighbors were playing loud music late into the night for several days in a row. I had enough of their loud music, so I called the cops on them every time they played their music. Now, my neighbors no longer play their loud music”*).

Some punishing actions were more creative, such as one participant who engineered a lesson for a roommate (*“My roommate kept leaving the broom in the way of the front door before leaving for work, so I arranged it in a way that it would fall when they came back and opened the door. This was done to remind them to put things back in the right place, all the time every*

time”), and another who bought termites for their revenge (“A person stole 3,000 from me, so I bought a termite colony and infested their house with it. I need more sentences, but that truly explains what happened. They stole from me, so I took matters in my own hand”).

Participants reported punishing both close others, like romantic partners (“I drive my husband to work every day, because he doesn't drive. I refused to drive him for a couple of days, because he ruined plans I had by working on an unscheduled day”), and strangers (“I was at the gas station when I gathered my items to purchase and I went to the counter to check out. I was a few feet away, next in line, when someone cut me and started to get checked out and I said to him wow so you're just going to cut me like that?”). Punishing actions took place in a range of settings, such as at home (“After cooking dinner for my partner, they refused to eat it because they didn't like how it tasted. They insulted my cooking skills. I didn't cook for my partner again until they apologized for their behavior. They apologized five days later”) or at work (“At work someone was repeatedly taking my bag of chips that I had with my lunch. When I found out who it was I punished them by not talking to them anymore and cutting off friendly conversations”).

Some participants explained why they avoided certain punitive actions (“My roommate sometimes annoys me but I can't confront him directly because it will lead to further problems. Instead, when he is asleep, I take his leftovers from the fridge and spit in it and then place it back into the fridge”), with this example also illustrating that some participants were willing to report behaviours that could be seen as vindictive or ‘gross’. Another example of a participant candidly describing undesirable behaviour was: “At the local Target someone took the parking spot that I was turning into, so I flicked them off and cursed at them.” Others focused on describing the impact of the violation and how they escalated punishment from warning to action (“Recently, I had a situation where a coworker consistently interrupted me during meetings, undermining my ability to communicate effectively and contribute to discussions. After attempting to address the issue politely multiple times without success, I brought it to the attention of our manager. As a result, the coworker received a formal warning and was required to attend communication skills training to improve their behavior”).

### ***Third-Party Punishing Behaviours***

Participants described a variety of third-party punishing actions, including verbal confrontation (“I chastised a man who let his dog poop on someone else's lawn and he didn't pick it up. I told him he was being rude and he should pick up after his dog”), physical confrontation (“I was



walking in the mall with a previous girlfriend. We were getting down the escalator when she told me that a man just groped her. I went up to the man and punched him in the chest really hard. We then both just walked away”), child discipline (“One of my children was being very rude to another, so I made her apologize and took away her phone for the day”), pettiness (“All day my sister, 14 years old, was being mean to my brother, who's 8 years old. I told her to stop multiple times but she didn't. As a snack I offered to make some toast, so I burnt her toast on purpose”), social withdrawal (“One of my friends agreed to purchase a clock from my boyfriend and pay for it when she got paid. She got the clock then didn't pay for it after numerous attempts to collect it. I stopped taking her calls and socializing with her as a result”), sabotage (“I noticed one of my co-workers, someone I generally had a low impression of, verbally haranguing another co-worker I considered my friend, who was a very nice and inoffensive person who did not deserve the treatment they were getting. As a result I held back some information the offending co-worker needed to do their job so they were late completing a project”), and punishing by appealing to authorities (“I reported someone to HR after hearing him make inappropriate comments about a female co-worker”). Some participants also explained why they chose one punishing tactic over another (“I was shopping with a group of friends. One member of our group was very condescending and rude to the person who was helping us check out. I didn't want to embarrass the employee any more than possible so I kept quiet inside the store. Since then, I have excluded this person from any social event I organize nor do I attend if I know they will be there”).

Participants punished actions of varying severity, from minor transgressions such as cutting in line (“Waiting in a long line at the grocery store, someone tried to cut in front of a customer a few spots ahead of me. I pointed out the end of the line”) to more serious violations such as theft (“While waiting for a train, I noticed a young woman acting somewhat strange. Multiple trains stopped at this station and she would rush into crowds as people boarded and disembarked, only to reappear elsewhere on the platform shortly thereafter. After one crowd, I noticed she now had a bag and was rifling through it. I pointed her out to security and explained what I had seen. When the officers approached her, she took off running. When my train finally made it and we were pulling out of the station, I could see her in the back of a police car while officers were bagging evidence”).

Participants provided various motives for punishing, such as upholding rules (“It is not a very serious issue but two of my cousins and I like to play Monopoly whenever they come over to my house. The problem is that they are very competitive and hate to lose to each other. A few days

*ago while we were playing, one of my cousins had to go use the bathroom so we took a break. I got up to go get a quick snack and on my way back I saw my other cousin move his token one step for whatever reason. I confronted him and told my cousin so he was kicked out of the game”), fairness (“During presentations at my office recently, a coworker had gotten up in front of the office and presented ideas that another coworker had come up without giving them any proper credit. The whole office loved the idea and made the stealing coworker look great. After the meeting, I stayed behind and let the manager know discreetly whose ideas those actually were and what had happened. It was not fair for someone to be passing off ideas that weren't theirs for the credit”), just deserts (“My cousin has been really mean to my Aunt (her mom) so I basically stopped talking to my cousin. She doesn't deserve people being nice to her”), holding others accountable and promoting behaviour change (“My friend didn't pay my other friend back. So I kept telling him he needs to, and annoying him about it until he did pay my other friend back. I also was trying to distance myself from him until he redeemed himself”), and fulfilling a job responsibility (“I work in local government for compliance in social services and discovered an instance of Medicaid fraud, robbing people with disabilities of crucial funding. I reported the fraudster to the federal authorities and their license to provide service was suspended”).*

Some participants provided extensive explanations, including descriptions of how their actions reflected their motives and traits (“I intervened after seeing a bunch of people bullying a stranger on public transit. I handled the matter gently but aggressively, fighting for the victim's right to safety and respect. My actions were motivated by a desire to fight injustice and safeguard individuals who were at risk of mistreatment. Personally, I see my intervention as an expression of empathy and dedication to creating a safe and inclusive atmosphere for all. Others may regard me as bold and principled for speaking up against wrongdoing and advocating for others' well-being, so generating a sense of solidarity and community”). Others emphasised that their actions were done for the right reasons (“After discovering that an acquaintance of a close friend was spreading false rumors about them, I spoke with others in their social circle about the behavior. I told them that I wasn't trying to spread gossip, but that I felt it was important they be held accountable for their actions”), while others focused on how the violation impacted their relationship with the violator (“I decided to confront a man who cheated on his girlfriend by sending explicit photos to other women. I stood up for his girlfriend and told him that his actions were harmful and that they have consequences. I decided to end my friendship with the man because of his lack of trustworthiness and commitment to his partner”).

# Appendix E

## Supplementary Information for Chapter 6

### E1 Hypotheses as Preregistered with *Communications Psychology*

| Question                                                                                                            | Hypothesis                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Sampling plan (e.g. power analysis)                                                                                                                                                                                                                                                                    | Analysis Plan                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Interpretation given to different outcomes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|---------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Primary Hypotheses                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| Q1. Are uncalculated decisions around the personal cost of helping / punishing used as a signal of trustworthiness? | <p>Helping:<br/>H1.1) Participants will be significantly less likely to check the cost of helping in the decision process observable condition than in the decision process hidden condition.</p> <p>Punishing:<br/>H1.2a) Participants will be significantly less likely to check the cost of punishment in the decision process observable condition than in the decision process hidden condition.</p> <p>H1.2b) Participants will make significantly faster punishing decisions in the decision process observable condition than in decision the process hidden condition.</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H1.1 &amp; H1.2a) The sample size for this model will be N = 1306 (653 Players A per condition in Experiment 1/2).</p> <p>H1.2b) The sample size for this model will be N = 1306 (653 Players A per condition in Experiment 3).</p> | <p>H1.1 &amp; H1.2a) We will run a logistic regression with checking decision (0 = did not check the cost, 1 = checked the cost) as a function of decision process observability (0 = process hidden, 1 = process observable).</p> <p>H1.2b) We will run a linear regression, predicting decision time as a function of decision process observability (0 = process hidden, 1 = process observable). If the amount of time spent deciding whether to punish is highly skewed, punishing decision time will be natural log transformed.</p> | <p>H1.1 &amp; H1.2a) A significant negative coefficient for observability (0 = decision process hidden, 1 = decision process observable) will be interpreted as evidence that participants are less likely to check the personal cost of helping/punishing when their decision process is observable compared to hidden (and therefore, that they are more likely to act uncalculatingly when their decision process can be observed). Otherwise, there is no evidence for H1.1/H1.2a.</p> <p>H1.2b) A significant negative coefficient for observability (0 = decision process hidden, 1 = decision process observable) will be interpreted as evidence that participants make faster punishing decisions (i.e., act uncalculatingly) when their decision process is</p> |

|                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|--------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | observable compared to hidden. Otherwise, there is no evidence for H1.2b.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Q2. Are uncalculated decisions around the personal cost of helping / punishing perceived as a signal of trustworthiness? | <p>Helping:<br/>H2.1) Observers will send significantly more of their endowment to helpers who did not check the personal cost of helping than to helpers who checked the cost.</p> <p>Punishing:<br/>H2.2a) Observers will send significantly more of their endowment to punishers who did not check the personal cost of punishing than to punishers who checked the cost.</p> <p>H2.2b) Observers will send significantly more of their endowment to punishers who made relatively fast (vs relatively slow) decisions to punish.</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H2.1 &amp; H2.2a) The sample size for this model will be N = 653 (Players B in the observable condition in Experiment 1/2).</p> <p>H2.2b) The sample size for this model will be N = 653 (Players B in the observable condition in Experiment 3).</p> | <p>Analyses will be restricted to the observable condition because Players B can only condition their trust on Player A decision processes in this condition. As Players B make two sending decisions (based on the two possible decisions made by Player A during the first stage), each sending decision will be treated as an observation and robust SEs will be clustered on observer ID to account for the non-independence of repeated observations from the same participant. The endowment sent will be transformed from pence sent to percentage of endowment sent for ease of interpretation.</p> <p>H2.1 &amp; H2.2a) We will run a linear regression predicting the percentage of endowment sent by Players B as a function of helpers/punishers cost-checking decisions (0 = did not check the personal cost, 1 = checked the cost).</p> | <p>H2.1 &amp; H2.2a) A significant negative coefficient for cost checking (0 = did not check the cost, 1 = checked the cost) will be interpreted as evidence that observers send a higher proportion of their endowment to helpers/punishers who did not check the cost of helping/ punishing compared to helpers/punishers who checked the cost of helping/punishing (i.e., that observers trust uncalculating helpers/punishers more than calculating helpers/punishers). Otherwise, there is no evidence for H2.1/H2.2a.</p> <p>H2.2b) A significant positive coefficient of decision speed (0 = relatively slow, 1 = relatively fast) will be interpreted as evidence that observers send a higher proportion of their endowment to punishers who decided relatively quickly compared to punishers who decided relatively slowly (i.e., that observers trust uncalculating punishers more than calculating punishers). Otherwise, there is no evidence for H2.2b.</p> |

|                                                                                                                                                                 |                                                                                                                                                                                 |                                                                                                                                                                                           |                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                      |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                                                                 |                                                                                                                                                                                 |                                                                                                                                                                                           | H2.2b) We will run a linear regression predicting the percentage of endowment Players B sent to punishers in the observable condition as a function of decision time (0 = relatively slow, 1 = relatively fast).                                                         |                                                                                                                                                                                                                                                                                                                                                                      |
| Q3. Does the operationalisation of uncalculating behaviour differentially influence the perceived trustworthiness of punishers in the context of personal cost? | H3) Observers will send significantly less of their endowment to punishers who check the personal cost of punishing than to punishers who take a long time to decide to punish. | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H3) The sample size for this model will be N = 1306 (Players B in the observable condition in Experiment 2 and 3).</p> | H3) We will run a linear regression predicting the percentage of endowment sent by Players B as a function of experiment (decision time vs cost checking) and deliberation (uncalculated vs calculated), as well as the interaction between experiment and deliberation. | H3) A significant negative coefficient for the interaction between experiment (0 = decision time, 1 = cost checking) and deliberation (0 = uncalculated, 1 = calculated) will be interpreted as evidence that the effect of calculated vs uncalculated punishment on trust is stronger for cost checking than decision time. Otherwise, there is no evidence for H3. |
| Q4. Do uncalculated helping and punishing decisions differentially influence perceived trustworthiness in the context of personal cost?                         | H4) Observers will send significantly less of their endowment to helpers who check the personal cost than to punishers who check the personal cost.                             | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H4) The sample size for this model will be N = 1306 (Players B in the observable condition in Experiment 1 and 2).</p> | H4) We will run a linear regression predicting the percentage of endowment sent by Players B as a function of behaviour (punishing vs helping) and deliberation (uncalculated vs calculated), as well as the interaction between behaviour and deliberation.             | H4) A significant negative coefficient for the interaction between behaviour (0 = punishing, 1 = helping) and deliberation (0 = uncalculated, 1 = calculated) will be interpreted as evidence that the effect of calculated vs uncalculated decisions on trust is stronger for helping than punishing. Otherwise, there is no evidence for H4.                       |
| Q5. Are uncalculated decisions around target impact used                                                                                                        | <p>Helping:</p> <p>H5.1) Participants will be significantly less likely to check the impact of helping in the</p>                                                               | Please refer to the Sampling plan in Methods for detail.                                                                                                                                  | H5.1 & H5.2) We will run a logistic regression with checking decision (0 = did not check the impact, 1 = checked the impact)                                                                                                                                             | H5.1) A significant negative coefficient for observability (0 = decision process hidden, 1 = decision process observable) will be interpreted as                                                                                                                                                                                                                     |

|                                                                                               |                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|-----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| as a signal of trustworthiness?                                                               | <p>decision process observable condition than in the decision process hidden condition.</p> <p>Punishing:<br/>H5.2) Participants will be significantly more likely to check the impact of punishment in the decision process observable condition than in the decision process hidden condition.</p> | H5.1 & H5.2) The sample size for this model will be N = 1306 (653 Players A per condition in Experiment 4/5).                                                                                     | as a function of decision process observability (0 = process hidden, 1 = process observable).                                                                                                                                                                                                                                                                                             | <p>evidence that participants are less likely to check the impact of helping when their decision process is observable compared to hidden (and therefore, that they are more likely to act uncalculatingly when their decision process can be observed). Otherwise, there is no evidence for H5.1.</p> <p>H5.2) A significant positive coefficient for observability (0 = decision process hidden, 1 = decision process observable) will be interpreted as evidence that participants are more likely to check the impact of punishing when their decision process is observable compared to hidden (and therefore, that they are more likely to act calculatingly when their decision process can be observed). Otherwise, there is no evidence for H5.2.</p> |
| Q6. Are uncalculated decisions around target impact perceived as a signal of trustworthiness? | <p>Helping:<br/>H6.1) Observers will send significantly more of their endowment to helpers who did not check the impact of helping on targets than to helpers who checked the impact.</p> <p>Punishing:<br/>H6.2) Observers will send significantly more of their endowment to punishers who</p>     | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H6.1 &amp; H6.2) The sample size for this model will be N = 653 (Players B in the observable condition in Experiment 4/5).</p> | Analyses will be restricted to the observable condition because Players B can only condition their trust on Player A decision processes in this condition. As Players B make two sending decisions (based on the two possible decisions made by Player A during the first stage), each sending decision will be treated as an observation and robust SEs will be clustered on observer ID | H6.1) A significant negative coefficient for impact checking (0 = did not check the impact, 1 = checked the impact) will be interpreted as evidence that observers send a higher proportion of their endowment to helpers who did not check the impact of helping compared to helpers who checked the impact of helping (i.e., that observers trust uncalculating helpers more than calculating helpers). Otherwise, there is no evidence for H6.1.                                                                                                                                                                                                                                                                                                            |

|                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                         | checked the impact of punishing on targets than to punishers who did not check the impact.                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                          | <p>to account for the non-independence of repeated observations from the same participant. The endowment sent will be transformed from pence sent to percentage of endowment sent for ease of interpretation.</p> <p>H6.1 &amp; H6.2) We will run a linear regression predicting the percentage of endowment sent by Players B as a function of helpers/punishers impact-checking decisions (0 = did not check the impact, 1 = checked the impact).</p> | <p>H6.2) A significant positive coefficient for impact checking (0 = did not check the impact, 1 = checked the impact) will be interpreted as evidence that observers send a higher proportion of their endowment to punishers who checked the impact of punishing compared to punishers who did not check the impact of punishing (i.e., that observers trust calculating punishers more than uncalculating punishers). Otherwise, there is no evidence for H6.2.</p>                                                   |
| Secondary Hypotheses                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Q7. Do uncalculated decisions around personal cost affect the perceived trustworthiness of non-helpers / non-punishers? | <p>Helping:<br/>H7.1) Observers will send significantly more of their endowment to non-helpers who checked the cost of helping than to non-helpers who did not check the cost of helping.</p> <p>Punishing:<br/>H7.2a) Observers will send significantly more of their endowment to non-punishers who did not check the cost of punishing than to non-punishers</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H7.1 &amp; H7.2a) The sample size for this model will be N = 653 (Players B in the observable condition in Experiment 1/2).</p> <p>H7.2b) The sample size for this model will be N = 653 (Players B in the observable condition in Experiment 3).</p> | <p>Analysis will be restricted to the observable condition because Players B can only condition their trust on Player A decision processes in this condition. As each observer makes two sending decisions, we will cluster robust SEs on observer ID.</p> <p>H7.1 &amp; H7.2a) We will run a linear regression predicting the percentage of endowment Players B sent to non-helpers/ non-punishers as a function of cost checking decisions.</p>       | <p>H7.1) A significant positive coefficient for cost checking (0 = did not check the cost, 1 = checked the cost) will be interpreted as evidence that observers send more of their endowment to non-helpers who checked the cost of helping compared to non-helpers who did not check the cost of helping (i.e., observers trust calculating non-helpers more than uncalculating non-helpers). Otherwise, there is no evidence for H7.1.</p> <p>H7.2a) A significant negative coefficient for cost checking (0 = did</p> |

|                                                                                                                                  |                                                                                                                                                                                           |                                                                                                                                                           |                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|----------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                                  | <p>who checked the cost of punishing.</p> <p>H7.2b) Observers will send significantly more of their endowment to relatively fast non-punishers than to relatively slow non-punishers.</p> |                                                                                                                                                           | <p>H7.2b) We will run a linear regression predicting the percentage of endowment Players B sent to non-punishers as a function of decision speed.</p>                                                                                                       | <p>not check the cost, 1 = checked the cost) will be interpreted as evidence that observers send more of their endowment to non-punishers who did not check the cost of punishing compared to non-punishers who checked the cost (i.e., observers trust uncalculating non-punishers more than calculating non-punishers). Otherwise, there is no evidence for H7.2a.</p> <p>H7.2b) A significant positive coefficient for decision speed (0 = relatively slow, 1 = relatively fast) will be interpreted as evidence that observers send more of their endowment to non-punishers who take little time in their decision not to punish compared to those who take a long time to decide not to punish (i.e., observers trust uncalculating non-punishers more than calculating non-punishers). Otherwise, there is no evidence for H7.2b.</p> |
| Q8. Do uncalculated decisions around personal cost have a stronger effect on the perceived trustworthiness of helpers/ punishers | <p>Helping:</p> <p>H8.1) Players B will send significantly less of their endowment to helpers who checked the cost of helping than to non-helpers who checked the cost of helping.</p>    | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H8.1 &amp; H8.2a) The sample size for this model will be N = 653 (Players B in the</p> | <p>H8.1 &amp; H8.2a) We will run a linear regression predicting the percentage of endowment sent by Players B as a function of helping/punishing decision, cost checking decision, and the interaction between the two. Robust SEs will be clustered on</p> | <p>H8.1 &amp; H8.2a) A significant negative coefficient for the interaction between helping/punishing (0 = did not help/punish, 1 = helped/punished) and cost checking decisions (0 = did not check the cost, 1 = checked the cost) will be interpreted as evidence that the effect of uncalculating</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |



|                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|----------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| than non-helpers / non-punishers?                                                                                                                        | <p>Punishing:<br/>H8.2a) Players B will send significantly less of their endowment to punishers who checked the cost of punishing than to non-punishers who checked the cost of punishing.</p> <p>H8.2b) Players B will send significantly less of their endowment to relatively slow punishers than to relatively slow non-punishers.</p>             | <p>observable condition in Experiment 1/2).</p> <p>H8.2b) The sample size for this model will be N = 653 (Players B in the observable condition in Experiment 3).</p>                                                                                                                                  | <p>observer ID, as Players B will make four sending decisions based on each of the four possible Player A choices.</p> <p>H8.2b) We will run a linear regression predicting the percentage of endowment sent by Players B as a function of punishing decision, decision speed and the interaction between the two. Robust SEs will be clustered on participant ID, accounting for repeated observations (four per participant).</p>                | <p>behaviour on trust is larger when Players A decided to help/ punish compared to when Players A decided not to help/punish. Otherwise, there is no evidence for H8.1/H8.2a.</p> <p>H8.2b) A significant negative interaction between punishing decision (0 = did not punish, 1 = punished) and decision speed (0 = relatively slow, 1 = relatively fast) will be interpreted as evidence that the effect of uncalculating behaviour on trust is larger when Players A decided to punish compared to when Player As decided not to punish. Otherwise, there is no evidence for H8.2b.</p> |
| Q9. Do uncalculated decisions around personal cost have a stronger effect on the trustworthiness of helpers/ punishers than non-helpers / non-punishers? | <p>Helping:<br/>H9.1) Helpers who checked the cost of helping will return significantly less of their endowment than non-helpers who checked the cost of helping.</p> <p>Punishing:<br/>H9.2a) Punishers who checked the cost of punishing will return significantly less of their endowment than non-punishers who checked the cost of punishing.</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H9.1 &amp; H9.2a) The sample size for this model will be N = 1306 (653 Players A per condition in Experiment 1/2).</p> <p>H9.2b) The sample size for this model will be N = 1306 (653 Players A per condition in Experiment 3).</p> | <p>H9.1 &amp; H9.2a) We will run a linear regression predicting the percentage of endowment returned by Players A as a function of helping/punishing decision, cost checking decision, as well as the interaction between the two.</p> <p>H9.2b) We will run a linear regression predicting the percentage of endowment returned by Players A as a function of punishing decision, log-transformed punishing decision time, their interaction,</p> | <p>H9.1 &amp; H9.2a) A significant negative coefficient for the interaction between helping/punishing (0 = did not help/punish, 1 = helped/punished) and cost checking decisions (0 = did not check the cost, 1 = checked the cost) will be interpreted as evidence that the effect of uncalculating decision making on trustworthiness is larger when Players A decide to help/punish compared to when Players A decide not to help/punish. Otherwise, there is no evidence for H9.1/H9.2a.</p>                                                                                           |

|                                                                                                                                                                      |                                                                                                                                                                             |                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                                                                      | H9.2b) Fast punishers will return significantly less of their endowment than fast non-punishers.                                                                            |                                                                                                                                                                                                | as well as log-transformed general comprehension speed. As the analysis is correlational, we wish to avoid concerns that the punishing decision time is reflective of general comprehension and reading speed rather than only of the time taken to consider whether to punish. Therefore, the natural log-transformed time spent reading the comprehension questions (i.e., the sum of time spent on the three comprehension question pages) will be included as a control for comprehension and reading speed. | H9.2b) A significant negative interaction between punishing decision (0 = did not punish, 1 = punished) and log-transformed decision time will be interpreted as evidence that decision time is a stronger predictor of untrustworthiness when Player A punished versus did not punish. Otherwise, there is no evidence for H9.2b.                                                                       |
| Q10. Does the operationalisation of uncalculating behaviour differentially influence the perceived trustworthiness of non-punishers in the context of personal cost? | H10) Observers will send significantly less of their endowment to non-punishers who check the personal cost of punishing than non-punishers who take a long time to decide. | Please refer to the Sampling plan in Methods for detail.<br><br>H10) The sample size for this model will be N = 1306 (Players B in the observable condition in Experiment 2 and Experiment 3). | H10) We will run a linear regression predicting the percentage of endowment sent by Players B as a function of experiment (decision time vs cost checking) and deliberation (uncalculated vs calculated), as well as the interaction between experiment and deliberation.                                                                                                                                                                                                                                        | H10) A significant negative coefficient for the interaction between experiment (0 = decision time, 1 = cost checking) and deliberation (0 = uncalculated non-punishment, 1 = calculated non-punishment) will be interpreted as evidence that the effect of calculated vs uncalculated non-punishment on trust is stronger for cost checking than decision time. Otherwise, there is no evidence for H10. |
| Q11. Do uncalculated decisions around                                                                                                                                | Helping:<br>H11.1) Observers will send significantly more of their                                                                                                          | Please refer to the Sampling plan in Methods for detail.                                                                                                                                       | Analysis will be restricted to the observable condition because Players B can only condition their                                                                                                                                                                                                                                                                                                                                                                                                               | H11.1 & H11.2) A significant positive coefficient for impact checking (0 = did not check the impact, 1 = checked the                                                                                                                                                                                                                                                                                     |

|                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| target impact affect the perceived trustworthiness of non-helpers / non-punishers?                                                                                  | <p>endowment to non-helpers who checked the impact of helping than to non-helpers who did not check the impact.</p> <p>Punishing:<br/>H11.2) Observers will send significantly more of their endowment to non-punishers who checked the impact of punishing than to non-punishers who did not check the impact.</p>                                                                         | H11.1 & H11.2) The sample size for this model will be N = 653 (Players B in the observable condition in Experiment 4/5).                                                                            | <p>trust on Player A decision processes in this condition. As each observer makes two sending decisions, we will cluster robust SEs on observer ID.</p> <p>H11.1 &amp; H11.2) We will run a linear regression predicting the percentage of endowment Player Bs sent to non-helpers/ non-punishers as a function of impact checking decisions.</p>             | impact) will be interpreted as evidence that observers send more of their endowment to non-helpers/non-punishers who checked the impact of helping/punishing compared to non-helpers/non-punishers who did not check the impact of helping/punishing (i.e., observers trust calculating non-helpers/non-punishers more than uncalculating non-helpers/non-punishers). Otherwise, there is no evidence for H11.1/H11.2.                                                          |
| Q12. Do uncalculated decisions around target impact have a stronger effect on the perceived trustworthiness of helpers/ punishers than non-helpers / non-punishers? | <p>Helping:<br/>H12.1) Observers will send significantly less of their endowment to helpers who checked the impact of helping than to non-helpers who checked the impact of helping.</p> <p>Punishing:<br/>H12.2) Observers will send significantly less of their endowment to punishers who checked the impact of punishing than to non-punishers who checked the impact of punishing.</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H12.1 &amp; H12.2) The sample size for this model will be N = 653 (Players B in the observable condition in Experiment 4/5).</p> | H12.1 & H12.2) We will run a linear regression predicting the percentage of endowment sent by Players B as a function of helping/punishing decision, impact checking decision, and the interaction between the two. Robust SEs will be clustered on observer ID, as Players B will make sending decision based on each of the four possible Player A choices. | H12.1 & H12.2) A significant negative coefficient for the interaction between helping/punishing (0 = did not help/punish, 1 = helped/punished) and impact checking decisions (0 = did not check the impact, 1 = checked the impact) will be interpreted as evidence that the effect of uncalculating behaviour on trust is larger when Players A decided to help/punish compared to when Players A decided not to help/punish. Otherwise, there is no evidence for H12.1/H12.2. |
| Q13. Do uncalculated decisions around target impact have a stronger effect on the actual                                                                            | <p>Helping:<br/>H13.1) Helpers who checked the impact of helping will return significantly less of their endowment than non-helpers</p>                                                                                                                                                                                                                                                     | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H13.1 &amp; H13.2) The sample size for this model</p>                                                                            | H13.1 & H13.2) We will run a linear regression predicting the percentage of endowment returned by Players A as a function of helping/punishing decision, impact checking                                                                                                                                                                                      | H13.1 & H13.2) A significant negative coefficient for the interaction between helping/punishing (0 = did not help/punish, 1 = helped/punished) and impact checking decisions (0 = did not check the impact, 1 = checked the                                                                                                                                                                                                                                                     |

|                                                                                                                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| trustworthiness of helpers/ punishers than non-helpers / non-punishers?                                                                                                                                         | <p>who checked the impact of helping.</p> <p>Punishing:<br/>H13.2) Punishers who checked the impact of punishing will return significantly less of their endowment than non-punishers who checked the impact of punishing.</p>                                                                                                                                                                                                                                                                                                         | will be N = 1306 (653 Players A per condition in Experiment 4/5).                                                                                                                                                                                                                                                                                                                                                      | decision, as well as the interaction between the two.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | impact) will be interpreted as evidence that the effect of uncalculating decision making on trustworthiness is larger when Players A decide to help/punish compared to when Players A decide not to help/punish. Otherwise, there is no evidence for H13.1/H13.2.                                                                                                                                                                                                                                                                                                                                                                                                                      |
| <p>Preregistered Exploratory Hypotheses</p> <p>For all exploratory hypotheses, if power requirements are not achieved, the results will be reported as suggestive, pending confirmation in future research.</p> |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| Q14. Do uncalculated decisions around the personal cost of helping / punishing predict trustworthiness?                                                                                                         | <p>Helping:<br/>H14.1) Helpers who did not check the cost of helping will return significantly more of their endowment than helpers who checked the cost of helping.</p> <p>Punishing:<br/>H14.2a) Punishers who did not check the cost of punishing will return significantly more of their endowment than punishers who checked the cost of punishing.</p> <p>H14.2b) Punishers who made faster decisions to punish will return significantly more of their endowment than punishers who took a longer time to decide to punish.</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H14.1 &amp; H14.2a) As we do not know how many Player As will decide to help/punish, the sample size for this model will be up to N = 1306 (653 Player As per condition in Experiment 1/2).</p> <p>H14.2b) As we do not know how many Player As will decide to punish, the sample size for this model will be up to N = 1306 (653 Players A per</p> | <p>Here, both the observable and the hidden condition will be used, as we collect the data on Player A's decision process, even when Player B cannot observe it.</p> <p>H14.1 &amp; H14.2a) We will run a linear regression predicting the percentage returned as a function of the helpers/punishers cost checking decision.</p> <p>H14.2b) We will run a linear regression predicting the percentage of endowment returned by punishers as a function of log-transformed decision time and log-transformed general comprehension speed.</p> | <p>H14.1 &amp; H14.2a) A significant negative coefficient of cost checking (0 = did not check the cost, 1 = checked the cost) will be interpreted as evidence that helpers/punishers who do not check the personal cost before deciding to help/punish return more of their endowment than helpers/punishers who do check the cost (i.e., that uncalculating helpers/punishers are more trustworthy than calculating helpers/punishers). Otherwise, there is no evidence for H14.1/H14.2a.</p> <p>H14.2b) A significant negative coefficient of decision time will be interpreted as evidence that punishers who take a short time to make their decision to punish return more of</p> |

|                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|-----------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | condition in Experiment 3).                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                          | <p>their endowment than punishers who are slower in making their decision (i.e., that uncalculating punishers are more trustworthy than calculating punishers). Otherwise, there is no evidence for H14.2b.</p> <p>If power requirements are not achieved, the results will be reported as suggestive, pending confirmation in future research.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Q15. Do uncalculated decisions around personal cost predict the actual trustworthiness of non-helpers/ non-punishers? | <p>Helping:<br/>H15.1) Non-helpers who checked the cost of helping will return significantly more of their endowment than non-helpers who did not check the cost.</p> <p>Punishing:<br/>H15.2a) Non-punishers who did not check the cost of punishing will return significantly more of their endowment than non-punishers who checked the cost.</p> <p>H15.2b) Fast deciding non-punishers will return significantly more of their endowment than slow deciding non-punishers.</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H15.1 &amp; H15.2a) As we do not know how many Player As will decide not to help/punish, the sample size for this model will be N = 1306 (653 Players A per condition in Experiment 1/2).</p> <p>H15.2b) As we do not know how many Player As will decide not to punish, the sample size for this model will be N = 1306 (653 Players A per condition in Experiment 3).</p> | <p>H15.1 &amp; H15.2a) We will run a linear regression predicting the percentage of endowment returned by non-helpers/non-punishers as a function of cost checking behaviour.</p> <p>H15.2b) We will run a linear regression predicting the percentage of endowment returned by non-punishers as a function of log-transformed punishing decision time, controlling for log-transformed general comprehension speed.</p> | <p>H15.1) A significant positive coefficient for cost checking behaviour (0 = did not check the cost, 1 = checked the cost) will be interpreted as evidence that non-helpers who checked the cost of helping return more of their endowment compared to non-helpers who did not check the cost of helping (i.e., calculating non-helpers are more trustworthy than uncalculating non-helpers). Otherwise, there is no evidence for H15.1.</p> <p>H15.2a) A significant negative coefficient for cost checking behaviour (0 = did not check the cost, 1 = checked the cost) will be interpreted as evidence that non-punishers who did not check the cost of punishing return more of their endowment than non-punishers who checked the cost (i.e., uncalculating non-punishers are</p> |

|                                                                                                                                                               |                                                                                                                                                                    |                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                                                               |                                                                                                                                                                    |                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                               | <p>more trustworthy than calculating non-punishers). Otherwise, there is no evidence for H15.2a.</p> <p>H15.2b) A significant negative coefficient for log-transformed decision time will be interpreted as evidence that fast deciding non-punishers return more of their endowment than slow deciding non-punishers (i.e., uncalculating non-punishers are more trustworthy than calculating non-punishers). Otherwise, there is no evidence for H15.2b.</p> <p>If power requirements are not achieved, the results will be reported as suggestive, pending confirmation in future research.</p> |
| Q16. Does the operationalisation of uncalculating behaviour differentially influence the actual trustworthiness of punishers in the context of personal cost? | H16) Punishers who check the personal cost of punishing will return significantly less of their endowment than punishers who take a long time to decide to punish. | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H16) As we do not know how many Players A will decide to punish, the sample size for this model will be up to N = 2612 (653 Players A per condition in Experiments 2 and 3).</p> | H16) We will run a linear regression predicting the percentage of endowment returned by Players A as a function of experiment (decision time vs cost checking) and deliberation (uncalculated vs calculated), as well as the interaction between experiment and deliberation. | <p>H16) A significant negative coefficient for the interaction between experiment (0 = decision time, 1 = cost checking) and deliberation (0 = uncalculated, 1 = calculated) will be interpreted as evidence that the effect of calculated vs uncalculated punishment on trustworthiness is stronger for cost checking than decision time. Otherwise, there is no evidence for H16.</p> <p>If power requirements are not achieved, the results will be reported</p>                                                                                                                                |

|                                                                                                                                                                   |                                                                                                                                                                  |                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                                                                   |                                                                                                                                                                  |                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                               | as suggestive, pending confirmation in future research.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Q17. Does the operationalisation of uncalculating behaviour differentially influence the actual trustworthiness of non-punishers in the context of personal cost? | H17) Non-punishers who check the personal cost of punishing will return significantly less of their endowment than non-punishers who take a long time to decide. | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H17) As we do not know how many Player As will decide not to punish, the sample size for this model will be up to N = 2612 (653 Players A per condition in Experiments 2 and 32).</p> | H17) We will run a linear regression predicting the percentage of endowment returned by Players A as a function of experiment (decision time vs cost checking) and deliberation (uncalculated vs calculated), as well as the interaction between experiment and deliberation. | <p>H17) A significant negative coefficient for the interaction between experiment (0 = decision time, 1 = cost checking) and deliberation (0 = uncalculated non-punishment, 1 = calculated non-punishment) will be interpreted as evidence that the effect of calculated vs uncalculated non-punishment on trustworthiness is stronger for cost checking than decision time. Otherwise, there is no evidence for H17.</p> <p>If power requirements are not achieved, the results will be reported as suggestive, pending confirmation in future research.</p> |
| Q18. Do uncalculated helping and punishing decisions differentially influence actual trustworthiness in the context of personal cost?                             | H18) Helpers who check the personal cost of helping will return significantly less of their endowment than punishers who check the personal cost of punishing.   | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H18) As we do not know how many Players A will decide to help/punish, the sample size for this model will be up to N = 2612 (653 Players A per condition in Experiments 1 and 2).</p> | H18) We will run a linear regression predicting the percentage of endowment returned by Players A as a function of behaviour and deliberation, as well as the interaction between behaviour and deliberation.                                                                 | <p>H18) A significant negative coefficient for the interaction between behaviour (0 = punishing, 1 = helping) and deliberation (0 = uncalculated, 1 = calculated) will be interpreted as evidence that the effect of calculated vs uncalculated decisions on trustworthiness is stronger for helping than punishing. Otherwise, there is no evidence for H18.</p> <p>If power requirements are not achieved, the results will be reported</p>                                                                                                                 |

|                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|--------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                    | as suggestive, pending confirmation in future research.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Q19. Do uncalculated decisions around target impact predict the actual trustworthiness of helpers/punishers? | <p>Helping:<br/>H19.1) Helpers who did not check the impact of helping will return significantly more of their endowment than helpers who checked the impact of helping.</p> <p>Punishing:<br/>H19.2) Punishers who checked the impact of punishing will return significantly more of their endowment than punishers who did not check the impact of punishing.</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H19.1 &amp; H19.2) As we do not know how many Players A will decide to help/punish, the sample size for this model will be up to N = 1306 (653 Players A per condition in Experiment 4/5).</p> | <p>Here, both the observable and the hidden condition will be used, as we collect the data on Player A's decision process, even when Player B cannot observe it.</p> <p>H19.1 &amp; H19.2) We will run a linear regression predicting the percentage returned as a function of the helpers/punishers impact checking decision.</p> | <p>H19.1) A significant negative coefficient of impact checking (0 = did not check the impact, 1 = checked the impact) will be interpreted as evidence that helpers who do not check the impact on a target before deciding to help return more of their endowment than helpers who do check the impact (i.e., that uncalculating helpers are more trustworthy than calculating helpers). Otherwise, there is no evidence for H19.1.</p> <p>H19.2) A significant positive coefficient of impact checking (0 = did not check the impact, 1 = checked the impact) will be interpreted as evidence that punishers who do check the impact of punishing on targets before deciding to punish return more of their endowment than punishers who do not check the impact (i.e., that calculating punishers are more trustworthy than uncalculating punishers). Otherwise, there is no evidence for H19.2.</p> <p>If power requirements are not achieved, the results will be reported</p> |



|                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                 |                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|-----------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                 |                                                                                                                                                                                | as suggestive, pending confirmation in future research.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Q20. Do uncalculated decisions around target impact predict the actual trustworthiness of non-helpers/ non-punishers? | <p>Helping:<br/>H20.1) Non-helpers who checked the impact of helping will return significantly more of their endowment than non-helpers who did not check the impact.</p> <p>Punishing:<br/>H20.2) Non-punishers who checked the impact of punishing will return significantly more of their endowment than non-punishers who did not check the impact.</p> | <p>Please refer to the Sampling plan in Methods for detail.</p> <p>H20.1 &amp; H20.2) As we do not know how many Player As will decide not to punish/help, the sample size for this model will be N = 1306 (653 Players A per condition in Experiment 4/5).</p> | H20.1 & H20.2) We will run a linear regression predicting the percentage of endowment returned by non-helpers/non-punishers as a function of target impact checking behaviour. | <p>H20.1 &amp; H20.2) A significant positive coefficient for impact checking behaviour (0 = did not check the impact, 1 = checked the impact) will be interpreted as evidence that non-helpers/non-punishers who checked the impact of helping/punishing return more of their endowment compared to non-helpers/non-punishers who did not check the impact (i.e., calculating non-helpers/non-punishers are more trustworthy than uncalculating non-helpers/non-punishers). Otherwise, there is no evidence for H20.1/ H20.2.</p> <p>If power requirements are not achieved, the results will be reported as suggestive, pending confirmation in future research.</p> |

## E2 Selection of Study Materials

### Punishment Cost Checking (1.2a)

---

#### Start of Block: Information and consent

*This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, you will play one or two games. Your decisions during these games will affect your bonus payment and the bonus payment of other Prolific players.**

The study is estimated to last around X minutes.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. Only researchers working with the Raihani Lab will have access to datafiles containing identifying information (e.g. Prolific ID). Completely de-identified datafiles (e.g. no Prolific ID) will be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return your submission.

---

#### Local Data Protection Privacy Notice

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: 'Public task'

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return your submission.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)



Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can email the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
- (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
- (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).
- (5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning my submission.

- ☐ I **consent** to take part in this study. (1)
- ☐ I do **NOT consent** to take part in this study. (0)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent </strong>to take part in this study.*

---

Please enter your Prolific ID

---

#### Start of Block: Player A – Game 1

In addition to the base payment (£0.90) for participating in this study, you can earn a **bonus** based on your interaction with other Prolific participants during two different games.

Although some aspects of the two games are similar, in the first game you will be one of THREE players, and in the second game you will be one of TWO players.

---

#### **GAME 1**

GAME 1 is a three-player game. **You will be Player 3.**

**Player 1** and **Player 2** have already read all the rules and made their decisions, and are now waiting for yours.

In GAME 1, each player starts with the following amounts:

**Player 1: £0.10**

**Player 2: £0.00**

You: £0.10

Player 1 has to choose how much, if anything, to send to Player 2.

Any amount sent to Player 2 is tripled (e.g. for every 1p Player 1 sends, Player 2 receives 3p).

Player 2 then has to choose how much, if anything, to return to Player 1.

For instance, if Player 1 sends £0.10 to Player 2, Player 2 receives £0.30.

If Player 2 returns half, both players earn £0.15 and both have more than Player 1 started with.

If Player 2 returns nothing, Player 2 keeps £0.30 and Player 1 earns no bonus.

Players can also send or return any intermediate amount.

---

*Player 1 chose to send all £0.10 to Player 2, and therefore Player 2 received £0.30.*

*Player 2 chose to keep everything (£0.30) and to return nothing (£0.00).*

*As a result, Player 1 will now receive no bonus (£0.00) and Player 2 will receive a bonus of £0.30.*

You (Player 3) start this game with £0.10, and can pay a cost to punish Player 2 by removing some of their bonus. If you decide to punish, Player 2's bonus will be reduced by £0.15. (Player 1's bonus will not be affected by your decision.)

However, you DO NOT know how much the cost to yourself is, only that it is somewhere between £0.00 and £0.10.

If you choose to punish Player 2, you will keep whatever remains of your £0.10 bonus.

If you choose NOT to punish, you will keep all £0.10.

**Example:** If the cost to punish Player 2 turned out to be £0.05, you would keep £0.05.

**Example:** If the cost to punish Player 2 turned out to be £0.10, you would keep £0.00.

**Example:** If the cost to punish Player 2 turned out to be £0.00, you would keep £0.10.

While you currently do NOT know how much it would cost you to punish, you can request to find out before deciding whether to punish.

So, you must decide:

- 1) Whether or not to find out how much it would cost you to punish Player 2
- 2) Whether or not to punish Player 2

Please answer the following questions, to make sure you understand GAME 1. You **MUST** answer **ALL** questions correctly to receive your bonus!



Imagine that you are deciding whether or not to punish **Player 2**. Which *decision* that **you** make will likely result in the **highest** possible bonus for **you**?

- ☐ You deciding to punish (0)
- ☐ You deciding NOT to punish (1)
- ☐ Either decision could result in a higher bonus than the other decision, it solely depends on what the cost of punishing is (0)



Imagine that you are deciding whether or not to punish **Player 2**. Which *decision* that **you** make will result in the **lowest** possible bonus for **Player 2**?

- ☐ You deciding to punish (1)
- ☐ You deciding NOT to punish (0)
- ☐ Which decision results in the lowest bonus depends on what the cost of punishing is (0)



Imagine that you are deciding whether or not to punish **Player 2**. Which *decision* that **you** make will result in the **highest** possible bonus for **Player 1**?

- ☐ You deciding to punish (0)
- ☐ You deciding NOT to punish (0)
- ☐ Your decision does not affect Player 1's bonus (1)



What are the two choices that you have?

- ☐ Whether to find out the cost of punishing, and whether to punish Player 2 (1)
- ☐ Whether to punish Player 2 if the cost of punishing is large, and whether to punish Player 2 if the cost of punishing is small (0)
- ☐ Whether to punish Player 2, and how much to punish Player 2 (0)

---

Start of Block: Player A – Game 2

## **GAME 2**

After GAME 1, you will play GAME 2, which is a **two-player** game.

You will play GAME 2 with a NEW player who did NOT participate in GAME 1.

You will be the “**Receiver**”. The new player will be the “**Sender**”. There will be no further players.

In GAME 2:

- The **Sender** starts with £0.10
- The **Sender** chooses how much, if anything, to send to you (the **Receiver**)
- Any money sent to **you** is tripled: for every 1p the **Sender** sends, **you** receive 3p
- **You** then choose how much, if anything, to return to the **Sender**. You can return nothing, any intermediate amount, or everything.

If the **Sender** sends all £0.10, **you** will receive £0.30. If **you** return half of the £0.30, **you** and the **Sender** will both earn £0.15, and will both have more than the **Sender** started with.

But if **you** return £0.00, you will earn £0.30, while the **Sender** will earn £0.00.

*So, the **Sender** can gain or lose money by sending you money, depending on how much **you** return.*

Please answer the following questions, to make sure you understand GAME 2. **You MUST answer ALL questions correctly to receive your bonus!**

---



Imagine that the Sender is deciding how much to send to you. Which of the **Sender's** decisions will result in the **Sender** earning the highest payoff?

- ☐ The Sender deciding to send nothing (0)
  - ☐ The Sender deciding to send everything (0)
  - ☐ Which decision results in the highest payoff depends on how much you decide to return to the Sender (1)
- 



Imagine that you are deciding how much to return to the Sender. Which of **your** decisions will result in **you** earning the highest payoff?

- ☐ You deciding to return nothing (1)
  - ☐ You deciding to return everything (0)
  - ☐ Which decision results in the highest payoff depends on how much the Sender sent you (0)
- 



Did the **Sender** participate in GAME 1?

- ☐ Yes (0)
  - ☐ No (1)
- 

#### Start of Block: Player A Game 2 OBSERVABLE

Although the **Sender** did NOT participate in GAME 1, the **Sender** can base the decision of how much money to send to you on how **you** behaved in GAME 1.

**Specifically, the **Sender** can base this decision on whether **you** decided to punish **Player 2**, as well as whether **you** decided to find out the cost of punishing.**

Please answer the following question, to make sure you understand. You **MUST** answer ALL questions correctly to receive your bonus!

---



When deciding how much to send to **you**, what information can the **Sender** base this decision on?

- ☐ Nothing (0)
  - ☐ Whether you punished Player 2 (0)
  - ☐ Whether you decided to find out the cost of punishing (0)
  - ☐ Whether you punished Player 2, AND whether you decided to find out the cost of punishing (1)
- 



### **GAME 1**

Punishing entails **you** paying a cost to remove £0.15 from **Player 2's** £0.30 bonus. The cost is somewhere between £0.00 and £0.10, but you do not know exactly how much.

**NOTE: This study involves REAL PEOPLE.** You really do have the chance to find out the cost of punishing, and the Sender really does get to base their sending decision on whether you decide to find out.

**Would you like to find out the cost of punishing **Player 2**?**

- ☐ Yes (1)
  - ☐ No (0)
- 

Start of Block: Checked (Player A)

The cost of punishing is £0.05.

---



Right now **you** have a £0.10 bonus. You know that punishing would cost **you** £0.05 to reduce **Player 2's** £0.30 bonus by £0.15.



**Would you like to pay the cost to punish *Player 2*?**

☐ Yes (1)

☐ No (0)

---

Start of Block: Not Checked (Player A)

You decided not to find out what the cost of punishing is.



Right now **you** have a £0.10 bonus. Punishing means paying a cost to reduce **Player 2**'s £0.30 bonus by £0.15. The cost is somewhere between £0.00 and £0.10.

**Would you like to pay the cost to punish *Player 2*?**

☐ Yes (1)

☐ No (0)

---

Start of Block: Player A Game 2 HIDDEN

Although the **Sender** did NOT participate in GAME 1, the Sender can base the decision of how much money to send to you on how **you** behaved in GAME 1.

**Specifically, the **Sender** can base this decision on whether you decided to punish **Player 2** (but NOT on whether **you** decided to find out the cost of punishing).**

---



When deciding how much to send to you, what information can the **Sender** base this decision on?

- ☐ Nothing (0)
- ☐ Whether you punished Player 2 (1)
- ☐ Whether you decided to find out the cost of punishing (0)
- ☐ Whether you punished Player 2, AND whether you decided to find out the cost of punishing (0)
- 



### **GAME 1**

Punishing entails **you** paying a cost to remove £0.15 from **Player 2's** £0.30 bonus. The cost is somewhere between £0.00 and £0.10, but you do not know exactly how much.

**NOTE: This study involves REAL PEOPLE.** You really do have the chance to find out the cost of punishing.

**Would you like to find out the cost of punishing **Player 2**?**

- ☐ Yes (1)
- ☐ No (0)
- 

Start of Block: Player A Game Decisions

### **GAME 2**

The **Sender** will decide how much of their £0.10 to send to you.  
**You** will decide how much to return to the Sender.

Specifically, you will decide the percentage of pence you would like to return.

For example, if **you** decide to return 40% of the pence you receive...

- If the **Sender** sends £0.00 (and you receive £0.00) **you** will return £0.00
- If the **Sender** sends £0.10 (and you receive £0.30) **you** will return £0.12

**REMINDER: This study involves REAL PEOPLE**, and your decision will determine how much bonus you and the Sender actually receive.

***What percentage would you like to return?***

0 10 20 30 40 50 60 70 80 90 100

|                         |                                                                                    |
|-------------------------|------------------------------------------------------------------------------------|
| Percentage to return () |  |
|-------------------------|------------------------------------------------------------------------------------|

Start of Block: Player B – Game 1

In addition to the base payment (£0.90) for participating in this study, you can earn a **bonus** based on your interaction with other Prolific participants during two different games.

You will ONLY take part in the second game, and will NOT play in the first game. However, we would like you to read about and understand both games.

## **GAME 1**

GAME 1 is a three-player game.

Each player starts with the following amounts:

**Player 1: £0.10**

**Player 2: £0.00**

**Player 3: £0.10**

**Player 1** has to choose how much, if anything, to send to **Player 2**.

Any amount sent to **Player 2** is tripled (e.g. for every 1p **Player 1** sends, **Player 2** receives 3p).

**Player 2** then has to choose how much, if anything, to return to **Player 1**.

For instance, if **Player 1** sends £0.10 to Player 2, **Player 2** receives £0.30. If **Player 2** returns half, both players earn £0.15 and both have more than **Player 1** started with.

If **Player 2** returns nothing, then **Player 2** keeps £0.30 and **Player 1** earns no bonus.

Players can also send or return any intermediate amount.

**Player 1** and **Player 2** have already read all the rules and made their decisions.

## GAME 1

Player 1 chose to send all £0.10 to Player 2, and therefore Player 2 received £0.30.

Player 2 chose to keep everything (£0.30) and to return nothing (£0.00).

As a result, Player 1 will now receive no bonus (£0.00) and Player 2 will receive a bonus of £0.30.

Player 3 now has the chance to pay a cost to remove some of Player 2's bonus. If Player 3 punishes, Player 2's bonus will be reduced by £0.15. (Player 1's bonus will not be affected by Player 3's decision.)

Player 3 DOES NOT know how much this cost is, only that it is somewhere between £0.00 and £0.10.

If Player 3 chooses to punish Player 2 by paying the cost, Player 3 only gets to keep whatever remains of their initial £0.10 bonus.

If Player 3 chooses NOT to punish, Player 3 gets to keep all £0.10.

**Example:** If the cost to punish Player 2 turned out to be £0.05, Player 3 would keep £0.05.

**Example:** If the cost to punish Player 2 turned out to be £0.10, Player 3 would keep £0.00.

**Example:** If the cost to punish Player 2 turned out to be £0.00, Player 3 would keep £0.10.

While Player 3 initially does NOT know how much it costs them to punish, Player 3 can request to find out before deciding whether to punish.

So, Player 3 must decide:

- 1) Whether or not to find out how much it would cost them to punish Player 2
- 2) Whether or not to punish Player 2

Please answer the following questions, to make sure you understand GAME 1. **You MUST answer ALL questions correctly to receive your bonus!**



Imagine that Player 3 is deciding whether or not to punish Player 2. Which of Player 3's decisions will likely result in the **highest** possible bonus for Player 3?

- ☐ Player 3 deciding to punish (0)
- ☐ Player 3 deciding NOT to punish (1)
- ☐ Either decision could result in a higher bonus than the other decision, it solely depends on what the cost of punishing is (0)



Imagine that Player 3 is deciding whether or not to punish **Player 2**. Which of **Player 3's** decision will result in the **lowest** possible bonus for **Player 2**?

- ☐ Player 3 deciding to punish (1)
- ☐ Player 3 deciding NOT to punish (0)
- ☐ Which decision results in the lowest bonus depends on what the cost of punishing is (0)



Imagine that Player 3 is deciding whether or not to punish **Player 2**. Which of **Player 3's** decision will result in the **highest** possible bonus for **Player 1**?

- ☐ Player 3 deciding to punish (0)
- ☐ Player 3 deciding NOT to punish (0)
- ☐ Player 3's decision does not affect Player 1's bonus (1)



What are the two choices that **Player 3** has?

- ☐ Whether to find out the cost of punishing, and whether to punish Player 2 (1)
- ☐ Whether to punish Player 2 if the cost of punishing is large, and whether to punish Player 2 if the cost of punishing is small (0)
- ☐ Whether to punish Player 2, and how much to punish Player 2 (0)

---

Start of Block: Player B – Game 2

## **GAME 2**

GAME 2 is a two-player game. You will play this game with **Player 3** from GAME 1.

You will be the "**Sender**". **Player 3** will be the "**Receiver**".

In GAME 2:

- **You** (the Sender) start with £0.10
- **You** then choose how much, if anything, to send to the **Receiver**
- Any money **you** send to the **Receiver** is tripled: for every 1p **you** send, the **Receiver** receives 3p

- The **Receiver** then chooses how much, if anything, to return to you. The Receiver can return nothing, any intermediate amount, or everything.

If **you** send all £0.10, the **Receiver** will receive £0.30. If the **Receiver** returns half of the £0.30, **you** and the **Receiver** will both earn £0.15, and will both have more than **you** started with.

But if the **Receiver** returns nothing, they will earn £0.30, while **you** will earn nothing.

*So, **you** can gain or lose money by sending money to the Receiver, depending on how much the **Receiver** returns.*

Please answer the following questions, to make sure you understand GAME 2. **You MUST answer ALL questions correctly to receive your bonus!**



Imagine that you are deciding how much to send to the **Receiver**. Which of **your** decisions will result in the **highest** possible bonus for **you**?

- ☐ You deciding to send nothing (0)
- ☐ You deciding to send everything (0)
- ☐ Which decision results in the highest bonus depends on how much the Receiver decides to return to you (1)



Imagine that the Receiver is deciding how much to return to **you**. Which of the **Receiver's** decisions will result in the **highest** possible bonus for the **Receiver**?

- ☐ The Receiver deciding to return nothing (1)
  - ☐ The Receiver deciding to return everything (0)
  - ☐ Which decision results in the highest bonus depends on how much you sent to the Receiver (0)
-



Did the **Receiver** participate in GAME 1?

☐ Yes (1)

☐ No (0)

---

Start of Block: Player B Game 2 OBSERVABLE

### **GAME 2**

**You** currently have a £0.10 bonus.

**You** will now decide how much, if anything, to send to the **Receiver**.

**You** can base this decision on how the **Receiver** behaved as **Player 3** in GAME 1. Specifically, you can base your decision both on whether the **Receiver** decided to find out what the cost of punishing was, and whether the **Receiver** decided to punish **Player 2**.

In GAME 1, the cost of punishing turned out to be £0.05. So, punishing entailed **Player 3** sacrificing £0.05 to remove £0.15 from **Player 2**. But **Player 3** only knew what the cost of punishing was if they decided to find out.

In GAME 2, **you** will make **four different choices**, one for each of the possible **Player 3** decisions. **Your** choice that matches with the decision the **Receiver** actually made in GAME 1 will count for your bonus.

**NOTE: This study involves REAL PEOPLE, and your decisions will determine how much bonus you and the Receiver actually receive.**

---

Start of Block: Bdec1OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO punish **Player 2**, WITHOUT finding out how much it would cost them to punish?

0      2      4      6      8      10

Pence to send ()



---

Start of Block: Bdec2OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to punish **Player 2**, WITHOUT finding out how much it would cost them to punish?

0      2      4      6      8      10

|                  |                                                                                    |
|------------------|------------------------------------------------------------------------------------|
| Pence to send () |  |
|------------------|------------------------------------------------------------------------------------|

Start of Block: Bdec3OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO punish **Player 2**, WITH finding out how much it would cost them to punish?

0      2      4      6      8      10

|                  |                                                                                    |
|------------------|------------------------------------------------------------------------------------|
| Pence to send () |  |
|------------------|------------------------------------------------------------------------------------|

Start of Block: Bdec4OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to punish **Player 2**, WITH finding out how much it would cost them to punish?

0      2      4      6      8      10

|                  |                                                                                      |
|------------------|--------------------------------------------------------------------------------------|
| Pence to send () |  |
|------------------|--------------------------------------------------------------------------------------|

Start of Block: Player B Game 2 HIDDEN

## GAME 2

**You** currently have a £0.10 bonus.

**You** will now decide how much, if anything, to send to the **Receiver**.

**You** can base this decision on how the **Receiver** behaved as **Player 3** in GAME 1. Specifically, **you** can base your decision on whether the **Receiver** decided to punish **Player 2**.

In GAME 1, the cost of punishing turned out to be £0.05. So, punishing entailed **Player 3** sacrificing £0.05 to remove £0.15 from **Player 2**. But **Player 3** only knew what the cost of punishing was if they decided to find out.

In GAME 2, **you** will make **two different choices**, one for each of the possible **Player 3** decisions. **Your**



choice that matches with the decision the **Receiver** actually made in GAME 1 will count for your bonus.

*NOTE: This study involves REAL PEOPLE, and your decisions will determine how much bonus you and the Receiver actually receive.*

---

Start of Block: Bdec1hid

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO punish **Player 2**?

0      2      4      6      8      10

---

|                  |                                                                                    |
|------------------|------------------------------------------------------------------------------------|
| Pence to send () |  |
|------------------|------------------------------------------------------------------------------------|

---

Start of Block: Bdec2hid

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to punish **Player 2**?

0      2      4      6      8      10

---

|                  |                                                                                      |
|------------------|--------------------------------------------------------------------------------------|
| Pence to send () |  |
|------------------|--------------------------------------------------------------------------------------|

---

Start of Block: demographics

Thank you!

Lastly, please provide your demographic information.

-----

What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



What is your age in years?

---

---

**Start of Block: debrief**

**Debrief**

Player 3 / the Receiver and the Sender really are Prolific participants, and your decisions really do determine your bonus and the bonus of the other player. However, Player 1 and Player 2 in the first game are not real.

Please email [nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk) or message on Prolific if you would like further information.

*Thank you for your participation!*

Please advance to the next page to submit your responses.

---

## Punishment Personal Cost Decision Time Players A (1.2b)

---

### Start of Block: Information and consent

*This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, you will play one or two games. Your decisions during these games will affect your bonus payment and the bonus payment of other Prolific players.**

The study is estimated to last around X minutes.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. Only researchers working with the Raihani Lab will have access to datafiles containing identifying information (e.g. Prolific ID). Completely de-identified datafiles (e.g. no Prolific ID) will be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return your submission.

---

### Local Data Protection Privacy Notice

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: 'Public task'

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return your submission.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)

---



Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can email the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I

understand that according to data protection legislation, 'public task' will be the lawful basis for processing.

(3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.

(4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).

(5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning my submission.

☐ I **consent** to take part in this study. (1)

☐ I do **NOT consent** to take part in this study. (0)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent </strong>to take part in this study.*

Please enter your Prolific ID

---

### Start of Block: Game 1 comprehension

In addition to the base payment (£0.90) for participating in this study, you can earn a **bonus** based on your interaction with other Prolific participants during two different games.

Although some aspects of the two games are similar, in the first game you will be one of THREE players, and in the second game you will be one of TWO players.

---

### GAME 1

GAME 1 is a three-player game. **You will be Player 3.**

**Player 1** and **Player 2** have already read all the rules and made their decisions, and are now waiting for yours.

In GAME 1, each player starts with the following amounts:

**Player 1: £0.10**

**Player 2: £0.00**

**You: £0.10**

**Player 1** has to choose how much, if anything, to send to **Player 2**.

Any amount sent to **Player 2** is tripled (e.g. for every 1p **Player 1** sends, **Player 2** receives 3p).

**Player 2** then has to choose how much, if anything, to return to **Player 1**.

For instance, if **Player 1** sends £0.10 to Player 2, **Player 2** receives £0.30.  
If **Player 2** returns half, both players earn £0.15 and both have more than **Player 1** started with.  
If **Player 2** returns nothing, Player 2 keeps £0.30 and **Player 1** earns no bonus.

Players can also send or return any intermediate amount.

---

Timer1 Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

---

### GAME 1

**Player 1** chose to send all £0.10 to Player 2, and therefore **Player 2** received £0.30.

**Player 2** chose to keep everything (£0.30) and to return nothing (£0.00).

As a result, **Player 1** will now receive no bonus (£0.00) and **Player 2** will receive a bonus of £0.30.

**You (Player 3)** start this game with £0.10, and can pay a cost to punish **Player 2** by removing some of their bonus. If you decide to punish, **Player 2's bonus will be reduced by £0.15**. (**Player 1's** bonus will not be affected by your decision.)

However, **right now you do not know how much the cost to yourself is**, only that it is somewhere between £0.00 and £0.10

If **you** choose to punish **Player 2**, you will keep whatever remains of your £0.10 bonus.  
If you choose NOT to punish, you will keep all £0.10.

**Example:** If the cost to punish **Player 2** turned out to be £0.05, **you** would keep £0.05.

**Example:** If the cost to punish **Player 2** turned out to be £0.10, **you** would keep £0.00.

**Example:** If the cost to punish **Player 2** turned out to be £0.00, **you** would keep £0.10.

While you currently do not know how much it costs **you** to punish **Player 2**, we will tell you before you have to decide whether to punish.

Please answer the following questions, to make sure you understand GAME 1. **You MUST answer ALL questions correctly to receive your bonus!**

---



Imagine that you are deciding whether or not to punish **Player 2**. Which of *your decisions* will likely result in the **highest** possible bonus for *you*?

- ☐ You deciding to punish (0)
  - ☐ You deciding NOT to punish (1)
  - ☐ Either decision could result in a higher bonus than the other decision, it solely depends on what the cost of punishing is (0)
- 



Imagine that you are deciding whether or not to punish **Player 2**. Which of *your decisions* will result in the **lowest** possible bonus for **Player 2**?

- ☐ You deciding to punish (1)
  - ☐ You deciding NOT to punish (0)
  - ☐ Which decision results in the lowest bonus depends on what the cost of punishing is (0)
- 



Imagine that you are deciding whether or not to punish **Player 2**. Which of *your decisions* will result in the **highest** possible bonus for **Player 1**?

- ☐ You deciding to punish (0)
  - ☐ You deciding NOT to punish (0)
  - ☐ Your decision does not affect Player 1's bonus (1)
-



If you decide to punish **Player 2**, what is the cost to you?

- ☐ 0 pence (0)
- ☐ 5 pence (0)
- ☐ 10 pence (0)
- ☐ You do not know yet, but you will find out later (1)

---

### Start of Block: Game 2 comprehension

Timer2 Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

---

### **GAME 2**

After GAME 1, you will play GAME 2, which is a two-player game.

You will play GAME 2 with a NEW player who did NOT participate in GAME 1.

You will be the "**Receiver**". The new player will be the "**Sender**". There will be no further players.

In GAME 2:

- The **Sender** starts with £0.10
- The **Sender** chooses how much, if anything, to send to you (the **Receiver**)
- Any money sent to **you** is tripled: for every 1p the **Sender** sends, **you** receive 3p
- **You** then choose how much, if anything, to return to the **Sender**. You can return nothing, any intermediate amount, or everything.

If the **Sender** sends all £0.10, **you** will receive £0.30. If **you** return half of the £0.30, **you** and the **Sender** will both earn £0.15, and will both have more than the **Sender** started with.

But if **you** return £0.00, you will earn £0.30, while the **Sender** will earn £0.00.

So, the **Sender** can gain or lose money by sending you money, depending on how much you return.

Please answer the following questions, to make sure you understand GAME 2. **You MUST answer ALL questions correctly to receive your bonus!**



Imagine that the Sender is deciding how much to send to you. Which of the **Sender's** decisions will result in the **Sender** earning the highest payoff?

- ☐ The Sender deciding to send nothing (0)
- ☐ The Sender deciding to send everything (0)
- ☐ Which decision results in the highest payoff depends on how much you decide to return to the Sender (1)



Imagine that you are deciding how much to return to the Sender. Which of **your** decisions will result in **you** earning the highest payoff?

- ☐ You deciding to return nothing (1)
- ☐ You deciding to return everything (0)
- ☐ Which decision results in the highest payoff depends on how much the Sender sent you (0)



Did the **Sender** participate in GAME 1?

- ☐ Yes (0)
- ☐ No (1)

---

Start of Block: Observable



Timer3obs Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

---

Soon you will be told the cost of punishing. We will then record **how long it takes you to decide whether or not to punish Player 2.**

Although the **Sender** did NOT participate in GAME 1, they can base the decision of how much money to send to **you** in GAME 2 on how you behaved in GAME 1.

**Specifically, the Sender can base this decision on whether you decided to punish Player 2, and on how long it took you to decide whether to punish Player 2.**

Please answer the following question, to make sure you understand. You MUST answer ALL questions correctly to receive your bonus!

---

 When deciding how much to send to you, what information can the **Sender** base this decision on?

- ☐ Nothing (0)
- ☐ Whether you punished Player 2 (0)
- ☐ How long it took you to decide whether to punish Player 2 (0)
- ☐ Whether you punished Player 2, AND how long it took you to decide whether to punish Player 2 (1)

---

### **GAME 1**

On the next screen, we will tell you how much it costs to punish. **You** then decide whether to punish **Player 2.**

**NOTE: This study involves REAL PEOPLE.** In GAME 2, the **Sender** really does get to base their sending decision on whether **you** decide to punish in GAME 1, and on how long it takes you to decide whether to punish.

**We will record how long you spend on the next screen.**

---

Start of Block: Pun Decision

pun timer Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)



**Punishing costs £0.05.**

Punishing means sacrificing £0.05 of your £0.10 bonus to remove £0.15 from Player 2's £0.30 bonus.

**Would you like to punish **Player 2**?**

☐ Yes (1)

☐ No (0)

---

Start of Block: Hidden

Timer3hid Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

---

Soon you will be told the cost of punishing. We will then record **how long it takes you to decide whether or not to punish **Player 2**.**

Although the **Sender** did NOT participate in GAME 1, they can base the decision of how much money to send to you in GAME 2 on how **you** behaved in GAME 1.

**Specifically, the **Sender** can base this decision on whether **you** decided to punish **Player 2** (but NOT on how long it took **you** to decide whether to punish **Player 2**).**

Please answer the following question, to make sure you understand. *You MUST answer ALL questions correctly to receive your bonus!*



When deciding how much to send to you, what information can the **Sender** base this decision on?

- ☐ Nothing (0)
- ☐ Whether you punished Player 2 (1)
- ☐ How long it took you to decide whether to punish Player 2 (0)
- ☐ Whether you punished Player 2, AND how long it took you to decide whether to punish Player 2 (0)

---

### **GAME 1**

On the next screen, we will tell you how much it costs to punish. **You** then decide whether to punish **Player 2**.

**NOTE: This study involves REAL PEOPLE.** In GAME 2, the **Sender** really does get to base their sending decision on whether **you** decide to punish in GAME 1.

**We will record how long you spend on the next screen.**

---

Start of Block: Return Decision

### **GAME 2**

The **Sender** will decide how much of their £0.10 to send to you.  
**You** will decide how much to return to the Sender.

Specifically, you will decide the percentage of pence you would like to return.

For example, if **you** decide to return 40% of the pence you receive...

- If the **Sender** sends £0.00 (and you receive £0.00) **you** will return £0.00
- If the **Sender** sends £0.10 (and you receive £0.30) **you** will return £0.12

**REMINDER: This study involves REAL PEOPLE**, and your decision will determine how much bonus you and the Sender actually receive.

*What percentage would you like to return?*

0 10 20 30 40 50 60 70 80 90 100

|                          |                                                                                    |
|--------------------------|------------------------------------------------------------------------------------|
| Percentage to return (%) |  |
|--------------------------|------------------------------------------------------------------------------------|

Start of Block: demographics

Thank you!

Lastly, please provide your demographic information.

What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



What is your age in years?

Start of Block: debrief

## **Debrief**

The Sender really is another Prolific participant, and your decisions really do determine your bonus and the bonus of the other player. However, Player 1 and Player 2 in the first game are not real.

Please email [nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk) or message on Prolific if you would like further information.

*Thank you for your participation!*

Please advance to the next page to submit your responses.

---

## Punishment Personal Cost Decision Time Players B (1.2b)

---

### Start of Block: Information and consent

*This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, you will play one or two games. Your decisions during these games will affect your bonus payment and the bonus payment of other Prolific players.**

The study is estimated to last around X minutes.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. Only researchers working with the Raihani Lab will have access to datafiles containing identifying information (e.g. Prolific ID). Completely de-identified datafiles (e.g. no Prolific ID) will be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return your submission.

---

### Local Data Protection Privacy Notice

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: 'Public task'

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return your submission.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)

---



Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can email the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
- (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
- (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).
- (5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning my submission.

- ☐ I **consent** to take part in this study. (1)
- ☐ I do **NOT consent** to take part in this study. (0)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent </strong>to take part in this study.*

Please enter your Prolific ID

---

### Start of Block: Game 1 Instructions

In addition to the base payment (£0.90) for participating in this study, you can earn a **bonus** based on your interaction with other Prolific participants during two different games.

You will ONLY take part in the second game, and will NOT play in the first game. However, we would like you to read about and understand both games.

---

### **GAME 1**

GAME 1 is a three-player game.

Each player starts with the following amounts:

**Player 1: £0.10**

**Player 2: £0.00**

**Player 3: £0.10**

**Player 1** has to choose how much, if anything, to send to **Player 2**.

Any amount sent to **Player 2** is tripled (e.g. for every 1p **Player 1** sends, **Player 2** receives 3p).

**Player 2** then has to choose how much, if anything, to return to **Player 1**.

For instance, if **Player 1** sends £0.10 to Player 2, **Player 2** receives £0.30. If **Player 2** returns half, both players earn £0.15 and both have more than **Player 1** started with.

If **Player 2** returns nothing, then **Player 2** keeps £0.30 and **Player 1** earns no bonus. Players can also send or return any intermediate amount.

***Player 1** and **Player 2** have already read all the rules and made their decisions.*

---

### **GAME 1**

**Player 1** chose to send all £0.10 to Player 2, and therefore **Player 2** received £0.30.

**Player 2** chose to keep everything (£0.30) and to return nothing (£0.00).

As a result, **Player 1** will now receive no bonus (£0.00) and **Player 2** will receive a bonus of **£0.30**.

**Player 3** now has the chance to punish **Player 2** by removing some of their bonus. If **Player 3** punishes, **Player 2's** bonus will be reduced by £0.15. (**Player 1's** bonus will not be affected by Player 3's decision.)

At first, **Player 3** does not know how much this cost is, only that it is somewhere between £0.00 and £0.10.

***However, right before we ask **Player 3** to decide whether to punish **Player 2**, we tell **Player 3** that the cost is £0.05.*** Therefore, punishing entails **Player 3** sacrificing £0.05 of their £0.10 bonus to remove £0.15 from **Player 2's** £0.30 bonus.

After we reveal this information to Player 3, we **record how long **Player 3** takes to decide whether or not to punish **Player 2**.**

Please answer the following questions, to make sure you understand GAME 1. **You MUST answer ALL questions correctly to receive your bonus!**

---





Imagine that Player 3 is deciding whether or not to punish **Player 2**. Which of **Player 3's** decisions will result in the **highest** possible bonus for **Player 3**?

- ☐ Player 3 deciding to punish (0)
  - ☐ Player 3 deciding NOT to punish (1)
  - ☐ Either decision could result in a higher bonus than the other decision, it solely depends on what the cost of punishing is (0)
- 



Imagine that Player 3 is deciding whether or not to punish **Player 2**. Which of **Player 3's** decisions will result in the **lowest** possible bonus for **Player 2**?

- ☐ Player 3 deciding to punish (1)
  - ☐ Player 3 deciding NOT to punish (0)
  - ☐ Which decision results in the lowest bonus depends on what the cost of punishing is (0)
- 



Imagine that Player 3 is deciding whether or not to punish **Player 2**. Which of **Player 3's** decisions will result in the **highest** possible bonus for **Player 1**?

- ☐ Player 3 deciding to punish (0)
  - ☐ Player 3 deciding NOT to punish (0)
  - ☐ Player 3's decision does not affect Player 1's bonus (1)
- 



What does **Player 3** know about the cost of punishing **Player 2**?

- ☐ Player 3 never knows the exact cost of punishing (but does know that the cost is between 0p and 10p) (0)
- ☐ Player 3 always knows that the exact cost of punishing is 5p (0)
- ☐ Player 3 initially does not know the exact cost of punishing, but finds out that it is 5p right before deciding whether to punish Player 2 (1)

---

## Start of Block: Game 2 Instructions

### GAME 2

GAME 2 is a two-player game. You will play this game with **Player 3** from GAME 1.

You will be the "**Sender**". **Player 3** will be the "**Receiver**".

In GAME 2:

- **You** (the Sender) start with £0.10
- **You** then choose how much, if anything, to send to the Receiver
- Any money **you** send to the **Receiver** is tripled: for every 1p **you** send, the **Receiver** receives 3p
- The **Receiver** then chooses how much, if anything, to return to **you**. The Receiver can return nothing, any intermediate amount, or everything.

If **you** send all £0.10, the **Receiver** will receive £0.30. If the **Receiver** returns half of the £0.30, **you** and the **Receiver** will both earn £0.15, and will both have more than you started with.

But if the **Receiver** returns nothing, they will earn £0.30, while **you** will earn nothing.

*So, **you** can gain or lose money by sending money to the Receiver, depending on how much the **Receiver** returns.*

Please answer the following questions, to make sure you understand GAME 2. **You MUST answer ALL questions correctly to receive your bonus!**



Imagine that you are deciding how much to send to the **Receiver**. Which of **your decisions** will result in the **highest** possible bonus for **you**?

- ☐ You deciding to send nothing (0)
  - ☐ You deciding to send everything (0)
  - ☐ Which decision results in the highest bonus depends on how much the Receiver decides to return to you (1)
-



Imagine that the Receiver is deciding how much to return to you. Which of the **Receiver's** decisions will result in the **highest** possible bonus for the **Receiver**?

- ☐ The Receiver deciding to return nothing (1)
  - ☐ The Receiver deciding to return everything (0)
  - ☐ Which decision results in the highest bonus depends on how much you sent to the Receiver (0)
- 



Did the **Receiver** participate in GAME 1?

- ☐ Yes (1)
  - ☐ No (0)
- 

Start of Block: Observable

## **GAME 2**

Reminder: **Player 3** was told the cost of punishing is £0.05 right before deciding whether to punish **Player 2**, and then we recorded how long **Player 3** took to decide.

In GAME 1, half of **Player 3s** took X.XX seconds or longer to decide whether to punish. We will refer to these Player 3s as "taking a long time to decide". The other half of **Player 3s** took less than X.XX seconds to decide. We will refer to these Player 3s as "taking a short time to decide".

**You** currently have a £0.10 bonus. You will now decide how much, if anything, to send to the **Receiver**.

**You** can base this decision on how the **Receiver** behaved as **Player 3** in GAME 1.

**Specifically, you can base your decision on whether the Receiver decided to punish Player 2, and how long the Receiver took to decide.**

In GAME 2, **you** will make **four different choices**, one for each of the possible **Player 3** actions. **Your** choice that matches the action the **Receiver** actually took in GAME 1 will count for your bonus.

**NOTE: This study involves REAL PEOPLE, and your decisions will determine how much bonus you and the Receiver actually receive.**

---

Start of Block: dec1OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO punish **Player 2**, after taking a SHORT time to decide?

|                  |                                                                                    |   |   |   |   |    |
|------------------|------------------------------------------------------------------------------------|---|---|---|---|----|
|                  | 0                                                                                  | 2 | 4 | 6 | 8 | 10 |
| Pence to send () |  |   |   |   |   |    |

Start of Block: dec2OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to punish **Player 2**, after taking a SHORT time to decide?

|                  |                                                                                    |   |   |   |   |    |
|------------------|------------------------------------------------------------------------------------|---|---|---|---|----|
|                  | 0                                                                                  | 2 | 4 | 6 | 8 | 10 |
| Pence to send () |  |   |   |   |   |    |

Start of Block: dec3OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO punish **Player 2**, after taking a LONG time to decide?

|                  |                                                                                      |   |   |   |   |    |
|------------------|--------------------------------------------------------------------------------------|---|---|---|---|----|
|                  | 0                                                                                    | 2 | 4 | 6 | 8 | 10 |
| Pence to send () |  |   |   |   |   |    |

Start of Block: dec4OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to punish **Player 2**, after taking a LONG time to decide?

|                  |                                                                                      |   |   |   |   |    |
|------------------|--------------------------------------------------------------------------------------|---|---|---|---|----|
|                  | 0                                                                                    | 2 | 4 | 6 | 8 | 10 |
| Pence to send () |  |   |   |   |   |    |

Start of Block: Hidden

**GAME 2**

Reminder: **Player 3** was told the cost of punishing is £0.05 right before deciding whether to punish **Player 2**, and then we recorded how long **Player 3** took to decide.

In GAME 1, half of **Player 3s** took X.XX seconds or longer to decide whether to punish. We will refer to these Player 3s as "taking a long time to decide". The other half of **Player 3s** took less than X.XX seconds to decide. We will refer to these Player 3s as "taking a short time to decide".

**You** currently have a £0.10 bonus. You will now decide how much, if anything, to send to the **Receiver**.

**You** can base this decision on how the **Receiver** behaved as **Player 3** in GAME 1.

**Specifically, you can base your decision on whether the Receiver decided to punish Player 2.**

In GAME 2, **you** will make **two different choices**, one for each of the possible **Player 3** actions. Your choice that matches the action the **Receiver** actually took in GAME 1 will count for your bonus.

NOTE: This study involves REAL PEOPLE, and your decisions will determine how much bonus you and the Receiver actually receive.

---

Start of Block: dec1hid

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO punish **Player 2**?

|                  |                                                                                      |   |   |   |   |    |
|------------------|--------------------------------------------------------------------------------------|---|---|---|---|----|
|                  | 0                                                                                    | 2 | 4 | 6 | 8 | 10 |
| Pence to send () |  |   |   |   |   |    |

---

Start of Block: dec2hid

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to punish **Player 2**?

|                  |                                                                                      |   |   |   |   |    |
|------------------|--------------------------------------------------------------------------------------|---|---|---|---|----|
|                  | 0                                                                                    | 2 | 4 | 6 | 8 | 10 |
| Pence to send () |  |   |   |   |   |    |

---

Start of Block: demographics

Thank you!

Lastly, please provide your demographic information.

---

What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



What is your age in years?

\_\_\_\_\_

---

**Start of Block: debrief**

**Debrief**

Player 3 / the Receiver really was a Prolific participant, and your decisions really do determine your bonus and the bonus of the other player. However, Player 1 and Player 2 in the first game are not real.

Please email [nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk) or message on Prolific if you would like further information.

*Thank you for your participation!*

Please advance to the next page to submit your responses.

---

## Helping Impact Checking (2.1)

---

### Start of Block: Information and consent

*This study has been approved by the UCL Ethics Board. Project ID: ICN-PWB-NR-23-10-20A.*

The study is being run by Nicole Engeler (nicole.engeler.19@ucl.ac.uk) under the supervision of Professor Nichola Raihani (n.raihani@ucl.ac.uk) in the Department of Experimental Psychology at UCL.

**In this study, you will play one or two games. Your decisions during these games will affect your bonus payment and the bonus payment of other Prolific players.**

The study is estimated to last around X minutes.

All data will be handled according to the General Data Protection Regulations and will be kept anonymous. Only researchers working with the Raihani Lab will have access to datafiles containing identifying information (e.g. Prolific ID). Completely de-identified datafiles (e.g. no Prolific ID) will be shared online in accordance with open-science practices. If you do not consent to de-identified data being shared, please do not take part in this study. If you do not wish to complete this study, please return your submission.

---

### Local Data Protection Privacy Notice

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk)

The lawful basis that will be used to process your personal data is: 'Public task'

Your personal data will be processed so long as it is required for the research project. We will anonymise or pseudonymise the personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, or if you would like to contact us about your rights, please contact UCL in the first instance at [data-protection@ucl.ac.uk](mailto:data-protection@ucl.ac.uk).

**If you do not wish to participate, or if you change your mind during the course of the study, please return your submission.**

Ethics Officer: [ethics@ucl.ac.uk](mailto:ethics@ucl.ac.uk)

---



Please only consent to taking part in this study if you agree with ALL of the following statements:

- (1) I have read the information above and understand I can email the researchers with any questions.
- (2) I understand that my personal information will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
- (3) I understand that all personal information will remain confidential and that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
- (4) I understand that my anonymised research data may be shared with, and used by, others for future research (no one will be able to identify you when these data are shared).
- (5) I understand that I am free to withdraw from the study without penalty if I so wish, simply by returning my submission.

- ☐ I **consent** to take part in this study. (1)
- ☐ I do **NOT consent** to take part in this study. (0)

*Skip To: End of Survey If Please only consent to taking part in this study if you agree with ALL of the following statement... = I do <strong>NOT consent </strong>to take part in this study.*

Please enter your Prolific ID

---

### Start of Block: Player A – Game 1

In addition to the base payment (£0.90) for participating in this study, you can earn a **bonus** based on your interaction with other Prolific participants during two different games.

Although some aspects of the two games are similar, in the first game you will be one of THREE players, and in the second game you will be one of TWO players.

---

### **GAME 1**

GAME 1 is a three-player game. **You will be Player 3.**

**Player 1** and **Player 2** have already read all the rules and made their decisions, and are now waiting for yours.

In GAME 1, each player starts with the following amounts:

**Player 1: £0.10**



**Player 2: £0.00**

**You: £0.10**

**Player 1** has to choose how much, if anything, to send to **Player 2**.

Any amount sent to **Player 2** is tripled (e.g. for every 1p **Player 1** sends, **Player 2** receives 3p).

**Player 2** then has to choose how much, if anything, to return to **Player 1**.

For instance, if **Player 1** sends £0.10 to Player 2, **Player 2** receives £0.30.

If **Player 2** returns half, both players earn £0.15 and both have more than **Player 1** started with.

If **Player 2** returns nothing, Player 2 keeps £0.30 and **Player 1** earns no bonus.

Players can also send or return any intermediate amount.

---

### GAME 1

**Player 1** chose to send all £0.10 to Player 2, and therefore **Player 2** received £0.30.

**Player 2** chose to keep everything (£0.30) and to return nothing (£0.00).

As a result, **Player 1** will now receive no bonus (£0.00) and **Player 2** will receive a bonus of £0.30.

**You (Player 3)** start this game with £0.10, and can pay a cost to help **Player 1** by increasing their bonus. If you decide to help, the cost to yourself will be £0.05. (**Player 2's** bonus will not be affected by your decision.)

If you choose to help **Player 1**, **you** will keep £0.05. If you choose NOT to help, **you** will keep all £0.10.

However, **you DO NOT know how much helping will add to Player 1's bonus**, only that it is somewhere between £0.01 and £0.30.

**Example:** If the impact of helping turned out to be £0.01, **Player 1** would finish with £0.01.

**Example:** If the impact of helping turned out to be £0.30, **Player 1** would finish with £0.30.

While you currently do NOT know how much helping would impact **Player 1**, you can request to find out before deciding whether to help.

So, **you** must decide:

- 1) Whether or not to find out how much helping would impact **Player 1**
- 2) Whether or not to help **Player 1**

Please answer the following questions, to make sure you understand GAME 1. **You MUST answer ALL questions correctly to receive your bonus!**

---



Imagine that you are deciding whether or not to help **Player 1**. Which *decision* that **you** make will result in the **highest** possible bonus for **you**?

- ☐ You deciding to help (0)
  - ☐ You deciding NOT to help (1)
  - ☐ Either decision could result in a higher bonus than the other decision, it depends on what the impact of helping is (0)
- 



Imagine that you are deciding whether or not to help **Player 1**. Which *decision* that **you** make will result in the **highest** possible bonus for **Player 1**?

- ☐ You deciding to help (1)
  - ☐ You deciding NOT to help (0)
  - ☐ Which decision results in the highest bonus depends on what the impact of helping is (0)
- 



Imagine that you are deciding whether or not to help **Player 1**. Which *decision* that **you** make will result in the **lowest** possible bonus for **Player 2**?

- ☐ You deciding to help (0)
  - ☐ You deciding NOT to help (0)
  - ☐ Your decision does not affect Player 2's bonus (1)
-



What are the two choices that you have?

- ☐ Whether to find out the impact of helping, and whether to help Player 1 (1)
- ☐ Whether to help Player 1 if the impact of helping is large, and whether to help Player 1 if the impact of helping is small (0)
- ☐ Whether to help Player 1, and how much to help Player 1 (0)

---

Start of Block: Player A – Game 2

## **GAME 2**

After GAME 1, you will play GAME 2, which is a **two-player** game.

You will play GAME 2 with a NEW player who did NOT participate in GAME 1.

You will be the “**Receiver**”. The new player will be the “**Sender**”. There will be no further players.

In GAME 2:

- The **Sender** starts with £0.10
- The **Sender** chooses how much, if anything, to send to you (the **Receiver**)
- Any money sent to **you** is tripled: for every 1p the **Sender** sends, **you** receive 3p
- **You** then choose how much, if anything, to return to the **Sender**. You can return nothing, any intermediate amount, or everything.

If the **Sender** sends all £0.10, **you** will receive £0.30. If **you** return half of the £0.30, **you** and the **Sender** will both earn £0.15, and will both have more than the **Sender** started with.

But if **you** return £0.00, you will earn £0.30, while the **Sender** will earn £0.00.

*So, the **Sender** can gain or lose money by sending you money, depending on how much **you** return.*

Please answer the following questions, to make sure you understand GAME 2. **You MUST answer ALL questions correctly to receive your bonus!**

---



Imagine that the Sender is deciding how much to send to you. Which of the **Sender's** decisions will result in the **Sender** earning the highest payoff?

- ☐ The Sender deciding to send nothing (0)
  - ☐ The Sender deciding to send everything (0)
  - ☐ Which decision results in the highest payoff depends on how much you decide to return to the Sender (1)
- 



Imagine that you are deciding how much to return to the Sender. Which of **your** decisions will result in **you** earning the highest payoff?

- ☐ You deciding to return nothing (1)
  - ☐ You deciding to return everything (0)
  - ☐ Which decision results in the highest payoff depends on how much the Sender sent you (0)
- 



Did the **Sender** participate in GAME 1?

- ☐ Yes (0)
  - ☐ No (1)
- 

#### Start of Block: Player A Game 2 OBSERVABLE

Although the **Sender** did NOT participate in GAME 1, the **Sender** can base the decision of how much money to send to you on how **you** behaved in GAME 1.

**Specifically, the **Sender** can base this decision on whether **you** decided to help **Player 1**, as well as whether **you** decided to find out the impact of helping.**

Please answer the following question, to make sure you understand. You MUST answer ALL questions correctly to receive your bonus!

---



When deciding how much to send to **you**, what information can the **Sender** base this decision on?

- ☐ Nothing (0)
  - ☐ Whether you helped Player 1 (0)
  - ☐ Whether you decided to find out the impact of helping (0)
  - ☐ Whether you helped Player 1, AND whether you decided to find out the impact of helping (1)
- 

### **GAME 1**

Helping entails **you** paying £0.05 to increase **Player 1's** £0.00 bonus. The impact of helping is somewhere between £0.01 and £0.30, but you do not know exactly how much.

**NOTE: This study involves REAL PEOPLE.** *You really do have the chance to find out the impact of helping, and the Sender really does get to base their sending decision on whether you decide to find out.*

***Would you like to find out how much helping impacts **Player 1's** bonus?***

- ☐ Yes (1)
  - ☐ No (0)
- 

**Start of Block: Checked Impact (A)**

The impact of helping is £0.15.

---



Right now **you** have a £0.10 bonus. You know that helping would cost **you** £0.05 to increase **Player 1's** £0.00 bonus by £0.15.

**Would you like to pay the cost to help *Player 1*?**

☐ Yes (1)

☐ No (0)

---

Start of Block: NO look (A)

You decided not to find out what the impact of helping is.

-----



Right now **you** have a £0.10 bonus. You know that helping would cost **you** £0.05. Helping would add somewhere between £0.01 and £0.30 to **Player 1's** £0.00 bonus.

**Would you like to pay the cost to help *Player 1*?**

☐ Yes (1)

☐ No (0)

---

Start of Block: Player A Game 2 HIDDEN

Although the **Sender** did NOT participate in GAME 1, the Sender can base the decision of how much money to send to you on how **you** behaved in GAME 1.

**Specifically, the **Sender** can base this decision on whether **you** decided to help **Player 1** (but NOT on whether **you** decided to find out the impact of helping).**

-----



When deciding how much to send to you, what information can the **Sender** base this decision on?

- ☐ Nothing (0)
- ☐ Whether you helped Player 1 (1)
- ☐ Whether you decided to find out the impact of helping (0)
- ☐ Whether you helped Player 1, AND whether you decided to find out the impact of helping (0)



### **GAME 1**

Helping entails **you** paying £0.05 to increase **Player 1's** £0.00 bonus. The impact of helping is somewhere between £0.01 and £0.30, but you do not know exactly how much.

**NOTE: This study involves REAL PEOPLE.** *You really do have the chance to find out the impact of helping.*

**Would you like to find out how much helping impacts **Player 1's** bonus?**

- ☐ Yes (1)
- ☐ No (0)

---

Start of Block: Player A Game 2 Decisions

### **GAME 2**

The **Sender** will decide how much of their £0.10 to send to you.  
**You** will decide how much to return to the Sender.

Specifically, you will decide the percentage of pence you would like to return.

For example, if **you** decide to return 40% of the pence you receive...

- If the **Sender** sends £0.00 (and you receive £0.00) **you** will return £0.00
- If the **Sender** sends £0.10 (and you receive £0.30) **you** will return £0.12

**REMINDER: This study involves REAL PEOPLE,** and your decision will determine how much bonus

you and the Sender actually receive.

**What percentage would you like to return?**

0 10 20 30 40 50 60 70 80 90 100

|                         |                                                                                    |
|-------------------------|------------------------------------------------------------------------------------|
| Percentage to return () |  |
|-------------------------|------------------------------------------------------------------------------------|

---

### Start of Block: Player B – Game 1

In addition to the base payment (£0.90) for participating in this study, you can earn a **bonus** based on your interaction with other Prolific participants during two different games.

You will ONLY take part in the second game, and will NOT play in the first game. However, we would like you to read about and understand both games.

---

### GAME 1

GAME 1 is a three-player game.

Each player starts with the following amounts:

**Player 1: £0.10**

**Player 2: £0.00**

**Player 3: £0.10**

**Player 1** has to choose how much, if anything, to send to **Player 2**.

Any amount sent to **Player 2** is tripled (e.g. for every 1p **Player 1** sends, **Player 2** receives 3p).

**Player 2** then has to choose how much, if anything, to return to **Player 1**.

For instance, if **Player 1** sends £0.10 to Player 2, **Player 2** receives £0.30. If **Player 2** returns half, both players earn £0.15 and both have more than **Player 1** started with.

If **Player 2** returns nothing, then **Player 2** keeps £0.30 and **Player 1** earns no bonus.

Players can also send or return any intermediate amount.

***Player 1** and **Player 2** have already read all the rules and made their decisions.*

---

### GAME 1



Player 1 chose to send all £0.10 to Player 2, and therefore Player 2 received £0.30.

Player 2 chose to keep everything (£0.30) and to return nothing (£0.00).

As a result, Player 1 will now receive no bonus (£0.00) and Player 2 will receive a bonus of £0.30.

Player 3 now has the chance to pay a cost to help Player 1 by increasing their bonus. (Player 2's bonus will not be affected by Player 3's decision.)

If Player 3 helps, the cost to themselves will be £0.05. If Player 3 does NOT help, they will keep all £0.10.

However, Player 3 DOES NOT know how much helping will add to Player 1's bonus, only that it is somewhere between £0.01 and £0.30.

**Example:** If the impact of helping turned out to be £0.01, Player 1 would finish with £0.01.

**Example:** If the impact of helping turned out to be £0.30, Player 1 would finish with £0.30.

While Player 3 initially does NOT know how much help impacts Player 1, Player 3 can request to find out before deciding whether to help.

So, Player 3 must decide:

- 1) Whether or not to find out how much helping would impact Player 1
- 2) Whether or not to help Player 1

Please answer the following questions, to make sure you understand GAME 1. **You MUST answer ALL questions correctly to receive your bonus!**



Imagine that Player 3 is deciding whether or not to help Player 1. Which of Player 3's decisions will result in the **highest** possible bonus for Player 3?

- ☐ Player 3 deciding to help (0)
  - ☐ Player 3 deciding NOT to help (1)
  - ☐ Either decision could result in a higher bonus than the other decision, it depends on what the impact of helping is (0)
-



Imagine that Player 3 is deciding whether or not to help **Player 1**. Which of **Player 3's** decision will result in the **highest** possible bonus for **Player 1**?

- ☐ Player 3 deciding to help (1)
  - ☐ Player 3 deciding NOT to help (0)
  - ☐ Which decision results in the highest bonus depends on what the impact of helping is (0)
- 



Imagine that Player 3 is deciding whether or not to help **Player 1**. Which of **Player 3's** decision will result in the **lowest** possible bonus for **Player 2**?

- ☐ Player 3 deciding to help (0)
  - ☐ Player 3 deciding NOT to help (0)
  - ☐ Player 3's decision does not affect Player 2's bonus (1)
- 



What are the two choices that **Player 3** has?

- ☐ Whether to find out the impact of helping, and whether to help Player 1 (1)
  - ☐ Whether to help Player 1 if the impact of helping is large, and whether to help Player 1 if the impact of helping is small (0)
  - ☐ Whether to help Player 1, and how much to help Player 1 (0)
- 

Start of Block: Player B – Game 2

## **GAME 2**

GAME 2 is a two-player game. You will play this game with **Player 3** from GAME 1.

You will be the "**Sender**". **Player 3** will be the "**Receiver**".

In GAME 2:

- **You** (the Sender) start with £0.10
- **You** then choose how much, if anything, to send to the **Receiver**
- Any money **you** send to the **Receiver** is tripled: for every 1p **you** send, the **Receiver** receives 3p
- The **Receiver** then chooses how much, if anything, to return to you. The Receiver can return nothing, any intermediate amount, or everything.

If **you** send all £0.10, the **Receiver** will receive £0.30. If the **Receiver** returns half of the £0.30, **you** and the **Receiver** will both earn £0.15, and will both have more than **you** started with.

But if the **Receiver** returns nothing, they will earn £0.30, while **you** will earn nothing.

So, **you** can gain or lose money by sending money to the Receiver, depending on how much the **Receiver** returns.

Please answer the following questions, to make sure you understand GAME 2. **You MUST answer ALL questions correctly to receive your bonus!**



Imagine that you are deciding how much to send to the **Receiver**. Which of **your** decisions will result in the **highest** possible bonus for **you**?

- ☐ You deciding to send nothing (0)
- ☐ You deciding to send everything (0)
- ☐ Which decision results in the highest bonus depends on how much the Receiver decides to return to you (1)



Imagine that the Receiver is deciding how much to return to **you**. Which of the **Receiver's** decisions will result in the **highest** possible bonus for the **Receiver**?

- ☐ The Receiver deciding to return nothing (1)
- ☐ The Receiver deciding to return everything (0)
- ☐ Which decision results in the highest bonus depends on how much you sent to the Receiver (0)



Did the **Receiver** participate in GAME 1?

☐ Yes (1)

☐ No (0)

---

Start of Block: B G2 OBSERV

### **GAME 2**

**You** currently have a £0.10 bonus.

**You** will now decide how much, if anything, to send to the **Receiver**.

**You** can base this decision on how the **Receiver** behaved as **Player 3** in GAME 1. Specifically, **you** can base your decision both on whether the **Receiver** decided to find out what the impact of helping was, and whether the **Receiver** decided to help **Player 1**.

In GAME 1, the impact of helping turned out to be £0.15. So, helping entailed **Player 3** sacrificing £0.05 to give £0.15 to **Player 1**. But **Player 3** only knew what the impact of helping was if they decided to find out.

In GAME 2, **you** will make **four different choices**, one for each of the possible **Player 3** decisions. **Your** choice that matches with the decision the **Receiver** actually made in GAME 1 will count for your bonus.

**NOTE: This study involves REAL PEOPLE, and your decisions will determine how much bonus you and the Receiver actually receive.**

---

Start of Block: Bdec1OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO help, WITHOUT finding out how much helping would impact **Player 1**?

0      2      4      6      8      10

Pence to send ()



---

Start of Block: Bdec2OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to help, WITHOUT finding out how much helping would impact **Player 1**?

0      2      4      6      8      10

|                  |                                                                                    |
|------------------|------------------------------------------------------------------------------------|
| Pence to send () |  |
|------------------|------------------------------------------------------------------------------------|

Start of Block: Bdec3OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO help, WITH finding out how much helping would impact **Player 1**?

0      2      4      6      8      10

|                  |                                                                                    |
|------------------|------------------------------------------------------------------------------------|
| Pence to send () |  |
|------------------|------------------------------------------------------------------------------------|

Start of Block: Bdec4OBS

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to help, WITH finding out how much helping would impact **Player 1**?

0      2      4      6      8      10

|                  |                                                                                      |
|------------------|--------------------------------------------------------------------------------------|
| Pence to send () |  |
|------------------|--------------------------------------------------------------------------------------|

Start of Block: B G2 HIDDEN

## **GAME 2**

**You** currently have a £0.10 bonus.

**You** will now decide how much, if anything, to send to the **Receiver**.

**You** can base this decision on how the **Receiver** behaved as **Player 3** in GAME 1. Specifically, **you** can base your decision on whether the **Receiver** decided to help **Player 1**.

In GAME 1, the impact of helping turned out to be £0.15. So, helping entailed **Player 3** sacrificing

£0.05 to give £0.15 to **Player 1**. But **Player 3** only knew what the impact of helping was if they decided to find out.

In GAME 2, **you** will make **two different choices**, one for each of the possible **Player 3** decisions. **Your** choice that matches with the decision the **Receiver** actually made in GAME 1 will count for your bonus.

*NOTE: This study involves REAL PEOPLE, and your decisions will determine how much bonus you and the Receiver actually receive.*

---

Start of Block: Bdec1hid

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose TO help **Player 1**?

|                  |                                                                                      |   |   |   |   |    |
|------------------|--------------------------------------------------------------------------------------|---|---|---|---|----|
|                  | 0                                                                                    | 2 | 4 | 6 | 8 | 10 |
| Pence to send () |  |   |   |   |   |    |

---

Start of Block: Bdec2hid

How many pence would you like to send to a **Receiver** who, as **Player 3** in GAME 1...

Chose NOT to help **Player 1**?

|                  |                                                                                      |   |   |   |   |    |
|------------------|--------------------------------------------------------------------------------------|---|---|---|---|----|
|                  | 0                                                                                    | 2 | 4 | 6 | 8 | 10 |
| Pence to send () |  |   |   |   |   |    |

---

Start of Block: demographics

Thank you!

Lastly, please provide your demographic information.

---

What is your gender?

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3)
- ☐ Prefer not to say (4)



What is your age in years?

---

---

#### Start of Block: debrief

##### **Debrief**

Player 3 / the Receiver and the Sender really are Prolific participants, and your decisions really do determine your bonus and the bonus of the other player. However, Player 1 and Player 2 in the first game are not real.

Please email [nicole.engeler.19@ucl.ac.uk](mailto:nicole.engeler.19@ucl.ac.uk) or message on Prolific if you would like further information.

*Thank you for your participation!*

Please advance to the next page to submit your responses.

## E3 Bayes Factor Sensitivity Analysis

**Table E3**

### *Bayes Factor Sensitivity Analysis*

|        | All Participants |             |             | Excellent Comprehension |             |             |
|--------|------------------|-------------|-------------|-------------------------|-------------|-------------|
|        | Original Scale   | Scale x 0.5 | Scale x 1.5 | Original Scale          | Scale x 0.5 | Scale x 1.5 |
| H1.1   | <b>71.28</b>     | 54.86       | 62.34       | <b>309.86</b>           | 132.18      | 350.35      |
| H1.2a  | <b>6972.31</b>   | 3878.78     | 6754.62     | <b>1501.72</b>          | 588.75      | 1730.66     |
| H1.2b  | <b>2.16</b>      | 2.16        | 2.16        | <b>5.42</b>             | 5.42        | 5.42        |
| H2.1   | <b>6.00</b>      | 2.31        | 8.61        | <b>8.97</b>             | 2.38        | 22.42       |
| H2.2a  | <b>4.78</b>      | 1.89        | 5.49        | <b>0.72</b>             | 1.00        | 0.81        |
| H2.2b  | <b>0.39</b>      | 0.74        | 0.26        | <b>0.17</b>             | 0.34        | 0.12        |
| H3     | <b>0.10</b>      | 0.30        | 0.19        | <b>0.31</b>             | 0.54        | 0.21        |
| H4     | <b>140.10</b>    | 659.85      | 41.74       | <b>20.22</b>            | 17.30       | 17.07       |
| H5.1   | <b>3.86</b>      | 4.11        | 3.07        | <b>0.67</b>             | 0.91        | 0.51        |
| H5.2   | <b>0.60</b>      | 0.93        | 0.43        | <b>5.89</b>             | 4.73        | 5.28        |
| H6.1   | <b>1.53</b>      | 1.25        | 1.46        | <b>1.87</b>             | 1.28        | 2.39        |
| H6.2   | <b>4.31</b>      | 1.88        | 6.52        | <b>4.90</b>             | 1.70        | 10.11       |
| H7.1   | <b>0.68</b>      | 0.80        | 0.42        | <b>0.59</b>             | 0.81        | 0.46        |
| H7.2a  | <b>0.80</b>      | 0.87        | 0.56        | <b>1.20</b>             | 1.13        | 1.13        |
| H7.2b  | <b>0.26</b>      | 0.45        | 0.19        | <b>0.34</b>             | 0.58        | 0.23        |
| H8.1   | <b>1.16</b>      | 1.61        | 0.86        | <b>2.10</b>             | 2.22        | 1.73        |
| H8.2a  | <b>1.17</b>      | 1.57        | 0.84        | <b>0.66</b>             | 0.90        | 0.50        |
| H8.2b  | <b>0.28</b>      | 0.53        | 0.17        | <b>0.24</b>             | 0.43        | 0.16        |
| H9.1   | <b>18.29</b>     | 18.29       | 18.29       | <b>2.38</b>             | 2.38        | 2.38        |
| H9.2a  | <b>0.14</b>      | 0.14        | 0.14        | <b>0.17</b>             | 0.17        | 0.17        |
| H9.2b  | <b>0.37</b>      | 0.62        | 0.26        | <b>0.31</b>             | 0.54        | 0.22        |
| H10    | <b>0.75</b>      | 0.02        | 0.14        | <b>0.39</b>             | 0.62        | 0.26        |
| H11.1  | <b>5.05</b>      | 3.25        | 5.00        | <b>10.87</b>            | 4.04        | 15.00       |
| H11.2  | <b>0.48</b>      | 0.75        | 0.35        | <b>0.86</b>             | 1.00        | 0.73        |
| H12.1  | <b>2.13</b>      | 2.53        | 1.43        | <b>5.89</b>             | 4.83        | 5.07        |
| H12.2  | <b>0.79</b>      | 1.01        | 0.49        | <b>0.57</b>             | 0.80        | 0.42        |
| H13.1  | <b>8.59</b>      | 8.59        | 8.59        | <b>0.95</b>             | 0.95        | 0.95        |
| H13.2  | <b>0.13</b>      | 0.13        | 0.13        | <b>0.25</b>             | 0.25        | 0.25        |
| H14.1  | <b>97.33</b>     | 97.33       | 97.33       | <b>7746.67</b>          | 7746.67     | 7746.67     |
| H14.2a | <b>0.17</b>      | 0.17        | 0.17        | <b>0.31</b>             | 0.31        | 0.31        |
| H14.2b | <b>0.14</b>      | 0.14        | 0.14        | <b>0.49</b>             | 0.49        | 0.49        |
| H15.1  | <b>0.56</b>      | 0.56        | 0.56        | <b>0.27</b>             | 0.27        | 0.27        |
| H15.2a | <b>1.98</b>      | 1.98        | 1.98        | <b>1.03</b>             | 1.03        | 1.03        |
| H15.2b | <b>0.01</b>      | 0.01        | 0.01        | <b>0.03</b>             | 0.03        | 0.03        |
| H16    | <b>0.24</b>      | 0.24        | 0.24        | <b>0.16</b>             | 0.16        | 0.16        |
| H17    | <b>0.33</b>      | 0.33        | 0.33        | <b>0.14</b>             | 0.14        | 0.14        |
| H18    | <b>0.17</b>      | 0.17        | 0.17        | <b>0.48</b>             | 0.48        | 0.48        |
| H19.1  | <b>0.44</b>      | 0.44        | 0.44        | <b>0.62</b>             | 0.62        | 0.62        |
| H19.2  | <b>0.12</b>      | 0.12        | 0.12        | <b>0.24</b>             | 0.24        | 0.24        |
| H20.1  | <b>1.02</b>      | 1.02        | 1.02        | <b>0.44</b>             | 0.44        | 0.44        |
| H20.2  | <b>0.08</b>      | 0.08        | 0.08        | <b>0.12</b>             | 0.12        | 0.12        |
| Add. 1 | <b>0.98</b>      | 0.82        | 0.30        | <b>0.65</b>             | 0.84        | 0.42        |
| Add. 2 | <b>3.45</b>      | 3.45        | 3.45        | <b>0.46</b>             | 0.46        | 0.46        |

*Note.* This table presents Bayes factors for all hypotheses, both when all participants and only those with excellent comprehension are included. Bayes factors produced from the original scale are bolded, followed by results from scale adjustments (prior scale of 0.5 times and 1.5 times the original value).



# Bibliography

- Aarts, H., Gollwitzer, P. M., & Hassin, R. R. (2004). Goal contagion: perceiving is for pursuing. *Journal of personality and social psychology*, 87(1), 23.
- Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of personality and social psychology*, 93(5), 751.
- Abele, A. E., Cuddy, A. J., Judd, C. M., & Yzerbyt, V. Y. (2008). Fundamental dimensions of social judgment. *European Journal of Social Psychology*, 38(7), 1063-1065.
- Abu-Lughod, L. (1990). Can there be a feminist ethnography? *Women & Performance: a journal of feminist theory*, 5(1), 7-27.
- Adams, G. S., Mullen, E., & Guala, F. (2012). The social and psychological costs of punishing. *Behavioral and Brain Sciences*, 35(1), 15.
- Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health Psychology*, 19(6), 586-592.
- Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational research methods*, 17(4), 351-371.
- Aharoni, E., Simpson, D., Nahmias, E., & Gollwitzer, M. (2022). A painful message: Testing the effects of suffering and understanding on punishment judgments. *Zeitschrift für Psychologie*, 230(2), 138–151.
- Aiman-Smith, L., Scullen, S. E., & Barr, S. H. (2002). Conducting studies of decision making in organizational contexts: A tutorial for policy-capturing and other regression-based techniques. *Organizational Research Methods*, 5(4), 388-414.
- Albright, L., & Malloy, T. E. (2000). Experimental validity: Brunswik, Campbell, Cronbach, and enduring issues. *Review of general psychology*, 4(4), 337-353.
- Alcala, V., Johnson, K., Steele, C., Wu, J., Zhang, D., & Pashler, H. (2022). The tainted altruism effect: a successful pre-registered replication. *Royal Society open science*, 9(1), 211152.

Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public opinion quarterly*, 42(1), 93-104.

Alexander, R.D. (1979). *Darwinism and human affairs*. Seattle, WA: University of Washington Press.

Alexander, R.D. (1987). *The biology of moral systems*. New York, NY: Aldine de Gruyter.

Aloise-Young, P. A. (1993). The development of self-presentation: Self-promotion in 6-to 10-year-old children. *Social Cognition*, 11(2), 201-222.

Amemiya, J., Liu, Z., Compton, B. J., & Heyman, G. D. (2021). Children's judgements of positive claims people make about themselves. *Infant and Child Development*, 30(2), e2212.

Ames, D. R., & Wazlawek, A. S. (2014). Pushing in the dark: Causes and consequences of limited self-awareness for interpersonal assertiveness. *Personality and Social Psychology Bulletin*, 40(6), 775-790.

Amichai-Hamburger, Y., Wainapel, G., & Fox, S. (2002). " On the Internet no one knows I'm an introvert": Extroversion, neuroticism, and Internet interaction. *Cyberpsychology & behavior*, 5(2), 125-128.

Anderson, C., & Shirako, A. (2008). Are individuals' reputations related to their history of behavior? *Journal of personality and social psychology*, 94(2), 320.

Anderson, C., Ames, D. R., & Gosling, S. D. (2008). Punishing hubris: The perils of overestimating one's status in a group. *Personality and Social Psychology Bulletin*, 34(1), 90-101.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of personality and social psychology*, 9(3), 272.

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in cognitive sciences*, 24(9), 694-703.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, 100(401), 464-477.

Andreoni, J., & Gee, L. K. (2011). *Gun for hire: Does delegated enforcement crowd out peer punishment in giving to public goods?* (No. w17033). National Bureau of Economic Research.

- Andreoni, J., & Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics*, 88(7), 1605–1623.
- Archer, J., & Coyne, S. M. (2005). An Integrated Review of Indirect, Relational, and Social Aggression. *Personality and Social Psychology Review*, 9(3), 212–230.
- Argyris, C. (1975). Dangers in applying results from experimental social psychology. *American Psychologist*, 30(4), 469.
- Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American economic review*, 99(1), 544-555.
- Arseneau-Robar, T. J. M., Taucher, A. L., Müller, E., van Schaik, C., Bshary, R., & Willems, E. P. (2016). Female monkeys use both the carrot and the stick to promote male participation in intergroup fights. *Proceedings of the Royal Society B: Biological Sciences*, 283(1843), 20161817.
- Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology*, 6(3), 1614-2241.
- Bahry, D. L., & Wilson, R. K. (2006). Confusion or fairness in the field? Rejections in the ultimatum game under the strategy method. *Journal of Economic Behavior & Organization*, 60(1), 37–54.
- Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K. S., ... & Bian, L. (2015). *Psychological and sociomoral reasoning in infancy*. American Psychological Association.
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 0064.
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329-349.
- Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counterpunishment in one-shot interactions. *Economics Letters*, 122(2), 308-310.
- Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: a natural field experiment. *European Economic Review*, 56(8), 1773-1785.

- Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, 111(45), 15924-15927.
- Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, 7(1), 1-6.
- Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences*, 108(27), 11023-11027.
- Balliet, D., Molho, C., Columbus, S., & Cruz, T. D. D. (2022). Prosocial and punishment behaviours in everyday life. *Current opinion in psychology*, 43, 278-283.
- Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological bulletin*, 137(4), 594.
- Balliet, D., Tybur, J. M., & Van Lange, P. A. (2017). Functional interdependence theory: An evolutionary account of social situations. *Personality and Social Psychology Review*, 21(4), 361-388.
- Balliet, D., & Van Lange, P. A. (2013). Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on psychological science*, 8(4), 363-379.
- Banerjee, R. (2002). Children's understanding of self-presentational behavior: Links with mental-state reasoning and the attribution of embarrassment. *Merrill-Palmer Quarterly*, 48(4), 378-404.
- Banerjee, R., Bennett, M., & Luke, N. (2012). Children's reasoning about self-presentation following rule violations: The role of self-focused attention. *Child Development*, 83(5), 1805-1821.
- Banerjee, R., Heyman, G. D., & Lee, K. (2020). *The development of moral self-presentation*. In Jensen, L. A. (Ed.), *The Oxford handbook of moral development* (pp. 91-109). Oxford University Press.
- Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology*, 107(3), 393-413.

- Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the “tragedy of the commons”. *Evolution and Human Behavior*, 25(4), 209-220.
- Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, 27(5), 325-344.
- Barclay, P. (2010). Altruism as a courtship display: Some effects of third-party generosity on audience perceptions. *British Journal of Psychology*, 101(1), 123-135.
- Barclay, P. (2011). Competitive helping increases with the size of biological markets and invades defection. *Journal of theoretical biology*, 281(1), 47-55.
- Barclay, P. (2012). Harnessing the power of reputation: Strengths and limits for promoting cooperative behaviors. *Evolutionary Psychology*, 10(5), 147470491201000509.
- Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior*, 34(3), 164–175.
- Barclay, P. (2015). Reputation. *The handbook of evolutionary psychology*, 1-19.
- Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current opinion in psychology*, 7, 33-38.
- Barclay, P., & Barker, J.L. (2020). Greener than thou: people who protect the environment are more cooperative, compete to be environmental, and benefit from reputation. *Journal of Environmental Psychology*, 72, 101441.
- Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, 274(1610), 749-753.
- Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C. & Sugden, R. (2010). *Experimental Economics: Rethinking the Rules*. Princeton: Princeton University Press.
- Barker, J. L., Power, E. A., Heap, S., Puurtinen, M., & Sosis, R. (2019). Content, cost, and context: A framework for understanding human signaling systems. *Evolutionary Anthropology: Issues, News, and Reviews*, 28(2), 86-99.
- Barkow, J. H. (1992). Beneath new culture is old psychology: Gossip and social stratification. In Barkow, J. H., Tooby, J., & Cosmides, L. (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 627–637). Oxford University Press.

- Baron, J. (2001). Purposes and methods [Peer commentary on “Experimental practices in economics: A methodological challenge for psychologists?” by R. Hertwig & A. Ortmann]. *Behavioral and Brain Sciences*, 24(3), 383–451.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3), 255-278.
- Barranti, M., Carlson, E. N., & Furr, R. M. (2016). Disagreement about moral character is linked to interpersonal costs. *Social Psychological and Personality Science*, 7(8), 806-817.
- Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., ... & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, 113(17), 4688-4693.
- Bartley, T., & Child, C. (2014). Shaming the corporation: the social production of targets and the anti-sweatshop movement. *American Sociological Review*, 79, 653–679.
- Barton, K. (2020). Package ‘MuMIn’: Multi-Model Inference. R package version 1.46.0. <https://CRAN.R-project.org/package=MumIn>
- Bas-Hoogendam, J. M., van Steenbergen, H., van der Wee, N. J., & Westenberg, P. M. (2018). Not intended, still embarrassed: Social anxiety is related to increased levels of embarrassment in response to unintentional social norm violations. *European Psychiatry*, 52, 15-21.
- Bašić, Z., Falk, A., & Kosse, F. (2020). The development of egalitarian norm enforcement in childhood and adolescence. *Journal of Economic Behavior & Organization*, 179, 667-680.
- Bates, D. (2010). lme4: Mixed-effects modeling with R.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology letters*, 2(3), 412-414.
- Batistoni, T., Barclay, P., & Raihani, N. J. (2022). Third-party punishers do not compete to be chosen as partners in an experimental game. *Proceedings of the Royal Society B*, 289(1966), 20211773.

- Batson, C. D. (2011). *Altruism in humans*. Oxford University Press.
- Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, 40(2), 290–302.
- Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society*, 9(2), 171-192.
- Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1), 59–78.
- Baumard, N., & Sperber, D. (2010). Weird people, yes, but also weird experiments. *Behavioral and brain sciences*, 33(2-3).
- Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological bulletin*, 91(1), 3.
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of general psychology*, 5(4), 323-370.
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?. *Perspectives on psychological science*, 2(4), 396-403.
- Bazerman, M. H., & Chugh, D. (2006). Decisions without blinders. *Harvard business review*, 84(1), 88.
- Bedau, H. A., & Kelly, E. (2017). Punishment. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.
- Beer, A., & Watson, D. (2008). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment*, 90(3), 250-260.
- Beersma, B., & Van Kleef, G. A. (2012). Why people gossip: An empirical analysis of social motives, antecedents, and consequences. *Journal of Applied Social Psychology*, 42(11), 2640-2670.
- Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5), 1652–1678.

- Bereczkei, T., Birkas, B., & Kerekes, Z. (2007). Public charity offer as a proximate factor of evolved reputation-building strategy: An experimental analysis of a real-life situation. *Evolution and Human Behavior*, 28(4), 277–284.
- Berg, J., Dickhaut, J., & McCabe, K. A. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122–142.
- Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The Braggart's dilemma: On the social rewards and penalties of advertising prosocial behavior. *Journal of Marketing Research*, 52(1), 90-104.
- Berman, J. Z., & Silver, I. (2022). Prosocial behavior and reputation: When does doing good lead to looking good?. *Current opinion in psychology*, 43, 102-107.
- Bharucha, J. (2018). Creating an honest, transparent and productive workplace. *International Journal of Business Excellence*, 15(4), 467-482.
- Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate behavioral research*, 45(5), 853-885.
- Biesanz, J. C. (2021). The social accuracy model. *The Oxford handbook of accurate personality judgment*, 61-82.
- Biesanz, J. C., West, S. G., & Millevoi, A. (2007). What do you learn about someone over time? The relationship between length of acquaintance and consensus and self-other agreement in judgments of personality. *Journal of personality and social psychology*, 92(1), 119.
- Bliege Bird, R., & Power, E. A. (2015). Prosocial signaling and cooperation among Martu hunters. *Evolution and Human Behavior*, 36(5), 389-397.
- Bliege Bird, R., Ready, E., & Power, E. A. (2018). The social significance of subtle signals. *Nature human behaviour*, 2(7), 452-457.
- Bliege Bird, R., Scelza, B., Bird, D. W., & Smith, E. A. (2012). The hierarchy of virtue: Mutualism, altruism and signaling in Martu women's cooperative hunting. *Evolution and Human Behavior*, 33(1), 64–78.
- Boehm, C. (1993). Egalitarian Behavior and Reverse Dominance Hierarchy. *Current Anthropology*, 34(3), 227–254



- Boehm, C. (2001). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.
- Boehm, C. (2011). Retaliatory violence in human prehistory. *The British Journal of Criminology*, 51(3), 518-534.
- Böhm, R., & Regner, T. (2013). Charitable giving among females and males: an empirical test of the competitive altruism hypothesis. *Journal of Bioeconomics*, 15, 251-267.
- Boies, K., Fiset, J., & Gill, H. (2015). Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity. *The leadership quarterly*, 26(6), 1080-1094.
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual review of psychology*, 54(1), 579-616.
- Bolling, D. Z., Pitskel, N. B., Deen, B., Crowley, M. J., McPartland, J. C., Mayes, L. C., & Pelphrey, K. A. (2011). Dissociable brain mechanisms for processing social exclusion and rule violation. *NeuroImage*, 54(3), 2462–2471.
- Bolton, G. E., & Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic behavior*, 10(1), 95-121.
- Bone, J. E., & Raihani, N. J. (2015). Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior*, 36(4), 323-330.
- Bone, J. E., McAuliffe, K., & Raihani, N. J. (2016). Exploring the motivations for punishment: Framing and country-level effects. *PLoS One*, 11(8), e0159769.
- Boothby, E. J., Cooney, G., Sandstrom, G. M., & Clark, M. S. (2018). The liking gap in conversations: Do people like us more than we think?. *Psychological science*, 29(11), 1742-1756.
- Boyd, R. (2017). *A different kind of animal: How culture transformed our species*. Princeton University Press.
- Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6), 3531-3535.
- Boyd, R., & Richerson, P. J. (1989). The evolution of indirect reciprocity. *Social Networks*, 11(3), 213–236.

- Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, 13(3), 171–195.
- Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality?. *Proceedings of the Royal Society B: Biological Sciences*, 285(1875), 20180116.
- Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European journal of social psychology*, 41(2), 135-143.
- Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British journal of social psychology*, 51(1), 149-166.
- Brañas-Garza, P., Espín, A. M., Exadaktylos, F., & Herrmann, B. (2014). Fair and unfair punishers coexist in the Ultimatum Game. *Scientific reports*, 4(1), 6025.
- Brauer, M., & Chaurand, N. (2010). Descriptive norms, prescriptive norms, and social control: An intercultural comparison of people's reactions to uncivil behaviors. *European Journal of Social Psychology*, 40(3), 490-499.
- Brick, C., Fillon, A., Yeung, S. K., Wang, M., Lyu, H., Ho, J. Y. J., ... & Feldman, G. (2021). Self-interest is overestimated: Two successful pre-registered replications and extensions of Miller and Ratner (1998). *Collabra: Psychology*, 7(1), 23443.
- Brion, S., & Anderson, C. (2013). The loss of power: How illusions of alliance contribute to powerholders' downfall. *Organizational Behavior and Human Decision Processes*, 121(1), 129-139.
- Brion S., Lount R. B., Doyle S. P. (2015). Knowing if you are trusted: Does meta-accuracy promote trust development? *Social Psychological and Personality Science*, 6(7), 823–830.
- Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science*, 4(1), 2515245920960351.
- Brown, S., Gray, D., McHardy, J., & Taylor, K. (2015). Employee trust and workplace performance. *Journal of economic behavior & organization*, 116, 361-378.
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

- Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.
- Bshary, R., & Bergmüller, R. (2008). Distinguishing four fundamental approaches to the evolution of helping. *Journal of evolutionary biology*, 21(2), 405-420.
- Bshary, R., & Grutter, A. S. (2006). Image scoring and cooperation in a cleaner fish. *Nature*, 441, 975-978.
- Bshary, R., & Raihani, N. J. (2017). Helping in humans and other animals: A fruitful interdisciplinary dialogue. *Proceedings of the Royal Society B: Biological Sciences*, 284(1863), 20170929.
- Buckels, E. E., Jones, D. N., & Paulhus, D. L. (2013). Behavioral confirmation of everyday sadism. *Psychological science*, 24(11), 2201-2209.
- Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, 60(5), 930-940.
- Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., & Marois, R. (2015). From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron*, 87(6), 1369-1380.
- Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1-28.
- Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of wildlife management*, 912-918.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2), 261-304.
- Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral ecology and sociobiology*, 65, 23-35.
- Byron, K., & Landis, B. (2020). Relational misperceptions in the workplace: New frontiers and challenges. *Organization Science*, 31(1), 223-242.
- Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology*, 96(9), 2370-2382.

- Camden, C., Motley, M. T., & Wilson, A. (1984). White lies in interpersonal communication: A taxonomy and preliminary investigation of social motivations. *Western Journal of speech communication*, 48(4), 309-325.
- Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Russell Sage Foundation.
- Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, 19, 7-42.
- Cameron, J. J., & Vorauer, J. D. (2008). Feeling transparent: On metaperceptions and miscommunications. *Social and Personality Psychology Compass*, 2(2), 1093-1108.
- Cardenas, J. C. (2000). How do groups solve local commons dilemmas? Lessons from experimental economics in the field. *Environment, development and sustainability*, 2(3), 305-322.
- Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do I punish?: Deterrence and just deserts as motives for punishment. *Journal of personality and social psychology*, 83(2), 284.
- Carlson, E. N. (2016). Meta-accuracy and relationship quality: Weighing the costs and benefits of knowing what people really think about you. *Journal of Personality and Social Psychology*, 111(2), 250.
- Carlson, E. N., & Furr, R. M. (2009). Evidence of differential meta-accuracy: People understand the different impressions they make. *Psychological Science*, 20(8), 1033-1039.
- Carlson, E. N., & Kenny, D. A. (2012). Meta-accuracy: Do we know how others see us? In S. Vazire & T. D. Wilson (Eds.), *Handbook of selfknowledge* (pp. 242-257). Guilford Press.
- Carlson, E. N., Furr, R. M., & Vazire, S. (2010). Do we know the first impressions we make? Evidence for idiographic meta-accuracy and calibration of first impressions. *Social Psychological and Personality Science*, 1(1), 94-98.
- Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-insight: Do people really know how others see them?. *Journal of personality and social psychology*, 101(4), 831.
- Carlson, E. N., Vazire, S., & Oltmanns, T. F. (2011). You probably think this paper's about you: narcissists' perceptions of their personality and reputation. *Journal of personality and social psychology*, 101(1), 185.

- Carlson, R. W., & Zaki, J. (2018). Good deeds gone bad: Lay theories of altruism and selfishness. *Journal of Experimental Social Psychology*, 75, 36-40.
- Carpenter, J. P. (2003). Is fairness used instrumentally? Evidence from sequential bargaining. *Journal of Economic Psychology*, 24(4), 467-489.
- Carpenter, J., Kariv, S., & Schotter, A. (2012). Network architecture, cooperation and punishment in public good experiments. *Review of Economic Design*, 16, 93-118.
- Carroll, J. S., Perkowitz, W. T., Lurigio, A. J., & Weaver, F. M. (1987). Sentencing goals, causal attributions, ideology, and personality. *Journal of personality and social psychology*, 52(1), 107.
- Casari, M., & Luini, L. (2009). Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior & Organization*, 71(2), 273-282.
- Cavazza, N., Guidetti, M., & Pagliaro, S. (2015). Who cares for reputation? Individual differences and concern for reputation. *Current Psychology*, 34(1), 164-176.
- Cavazza, N., Pagliaro, S., & Guidetti, M. (2014). Antecedents of concern for personal reputation: The role of group entitativity and fear of social exclusion. *Basic and applied social psychology*, 36(4), 365-376.
- Champely, S. (2020). Pwr: basic functions for power analysis (R package version 1.3–0) [Computer software]. The Comprehensive R Archive Network. Retrieved from <https://CRAN.R-project.org/package=pwr>
- Charness, G., Cobo-Reyes, R., & Jiménez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior & Organization*, 68(1), 18-28.
- Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6), 1579-1601.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 158(3), 419-444.
- Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental economics*, 14, 47-83.
- Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, 39, 268-277.

- Chen, S., & Xu, Y. (2020). Warmth and competence: Impact of third-party punishment on punishers' reputation. *Acta Psychologica Sinica*, 52(12), 1436.
- Chetty, R., Hofmeyr, A., Kincaid, H., & Monroe, B. (2021). The trust game does not (only) measure trust: The risk-trust confound revisited. *Journal of Behavioral and Experimental Economics*, 90, 101520.
- Christensen R (2023). *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.1, <https://CRAN.R-project.org/package=ordinal>.
- Claessens, S., Atkinson, Q., & Raihani, N. (2024). Why do people punish? Evidence for a range of strategic concerns. Preprint.
- Clark, J. W., & Wink, K. (2012). The relationship between political ideology and punishment: What do jury panel members say?. *Applied Psychology in Criminal Justice*, 8(2).
- Clausell, E., & Fiske, S. T. (2005). When do subgroup parts add up to the stereotypic whole? Mixed stereotype content for gay male subgroups explains overall ratings. *Social Cognition*, 23(2), 161-181.
- Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, 373(6511), 209-216.
- Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An experimental ethnography. *Journal of personality and social psychology*, 70(5), 945.
- Cohen, J. E. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Columbus, S., Molho, C., Righetti, F., & Balliet, D. (2021). Interdependence and cooperation in daily life. *Journal of Personality and Social Psychology*, 120(3), 626.
- Combs, D. R., & Penn, D. L. (2004). The role of subclinical paranoia on social perception and behavior. *Schizophrenia research*, 69(1), 93-104.
- Connor, R. C. (1992). Egg-trading in simultaneous hermaphrodites: an alternative to Tit-for-Tat. *Journal of Evolutionary Biology*, 5, 523–528.
- Coricelli, G., Joffily, M., Montmarquette, C., & Villeval, M. C. (2010). Cheating, emotions, and rationality: an experiment on tax evasion. *Experimental Economics*, 13, 226-247.

- Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of personality and social psychology*, 92(2), 208.
- Crawley, M. J. (2013). *The R book*. Second Edition. Chichester: Wiley.
- Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, 4(3), 308-315.
- Crombag, H., Rassin, E., & Horselenberg, R. (2003). On vengeance. *Psychology, Crime and Law*, 9(4), 333-344.
- Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological bulletin*, 52(3), 177.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281.
- Cronk, L. (2007). The influence of cultural framing on play in the trust game: A Maasai example. *Evolution and Human Behavior*, 28(5), 352-358.
- Crowley, M. J., Wu, J., McCarty, E. R., David, D. H., Bailey, C. A., & Mayes, L. C. (2009). Exclusion and micro-rejection: Event-related potential response predicts mitigated distress. *Neuroreport*, 20(17), 1518-1522.
- Csikszentmihalyi, M., & Larson, R. (2014). *Flow and the foundations of positive psychology* (Vol. 10, pp. 978-94). Dordrecht: Springer.
- Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology*, 40, 61-149.
- Curry, O., & Dunbar, R. I. M. (2011). Altruism in networks: the effect of connections. *Biology letters*, 7(5), 651-653.
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353-380.
- Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass*, 10(2), 117-133.

- Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a “trembling hand” game. *PloS One*, 4(8), e6699.
- Cushman, F., Sarin, A., & Ho, M. (2019). Punishment as communication. *The Oxford handbook of moral psychology*, 197-209.
- Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, 17(12), 1082-1089.
- Dafoe, A., Renshon, J., & Huth, P. (2014). Reputation and status as motives for war. *Annual Review of Political Science*, 17, 371–393.
- Danielson, S. W., Stewart, E., & Vonasch, A. (2023). The Morality Map: Does living in a smaller community cause greater concern for moral reputation?. *Current Research in Ecological and Social Psychology*, 4, 100120.
- Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, 7(4), 324-336.
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1), 113.
- Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, 446(7137), 794-796.
- de Araujo, F. F. (2014). Do I look good in green? A conceptual framework integrating employee green behavior, impression management and social norms. *Amazônia, Organizações e Sustentabilidade*, 3(2), 7-23.
- De Cremer, D., & Tyler, T. R. (2005). Am I respected or not?: Inclusion and reputation as issues in group membership. *Social Justice Research*, 18(2), 121-153.
- de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., & van Dijk, E. (2019). Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology*, 84, 103800.
- Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, 38(6), 734-743.



- Dere, D., Zlomuzica, A., & Dere, E. (2019). Fellow travellers in cognitive evolution: Co-evolution of working memory and mental time travel?. *Neuroscience & Biobehavioral Reviews*, 105, 94-105.
- Desmoulins-Lebeault, F., Gajewski, J. F., & Meunier, L. (2018). Personality and risk aversion. *Economics Bulletin*, 38(1), 472-489.
- Deutchman, P., Bračić, M., Raihani, N., & McAuliffe, K. (2021). Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evolution and Human Behavior*, 42(1), 12-20.
- Dhaliwal, N. A., Patil, I., & Cushman, F. (2021). Reputational and cooperative benefits of third-party compensation. *Organizational Behavior and Human Decision Processes*, 164, 27-51.
- Dhaliwal, N. A., Skarlicki, D. P., Hoegg, J., & Daniels, M. A. (2020). Consequentialist motives for punishment signal trustworthiness. *Journal of Business Ethics*, 1-16.
- Dik, G., & Aarts, H. (2007). Behavioral cues to others' motivation and goal pursuits: The perception of effort facilitates goal inference and contagion. *Journal of Experimental Social Psychology*, 43(5), 727-737.
- Dimitroff, S. J., Harrod, E. G., Smith, K. E., Faig, K. E., Decety, J., & Norman, G. J. (2020). Third-party punishment following observed social rejection. *Emotion*, 20(4), 713.
- Ditto, P. H., & Jemmott, J. B. (1989). From rarity to evaluative extremity: effects of prevalence information on evaluations of positive and negative characteristics. *Journal of Personality and Social Psychology*, 57(1), 16.
- Dochtermann, N. A., & Jenkins, S. H. (2011). Developing multiple hypotheses in behavioral ecology. *Behavioral Ecology and Sociobiology*, 65, 37-45.
- Dores Cruz, T. D., Thielmann, I., Columbus, S., Molho, C., Wu, J., Righetti, F., & Balliet, D. (2021). Gossip and reputation in everyday life. *Philosophical Transactions of the Royal Society B*, 376(1838), 20200301.
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46.
- dos Santos, M. D., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, 278(1704), 371-377.

- dos Santos, M. D., Rankin, D. J., & Wedekind, C. (2013). Human cooperation based on punishment reputation. *Evolution*, 67(8), 2446-2450.
- dos Santos, M., & Wedekind, C. (2015). Reputation based on punishment rather than generosity allows for evolution of cooperation in sizable groups. *Evolution and Human Behavior*, 36(1), 59-64.
- Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1), 45-70.
- Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, 452(7185), 348-351.
- Dubey, I., Groom, M., Tahir, A., & Hamilton, A. (2024). Social anxiety is more likely to influence reputation management decisions than autistic traits. *Autism Research*, 17(1), 78-88.
- Duca, S., & Nax, H. H. (2018). Groups and scores: the decline of cooperation. *Journal of The Royal Society Interface*, 15(144), 20180158.
- Dukerich, J.M., & Carter, S.M. (2000). Distorted images and reputation repair. In M. Schultz, M.J. Hatch, & M.H. Larsen (Eds.), *The expressive organization: linking identity, reputation, and the corporate brand* (pp. 97–112). New York, NY: Oxford University Press.
- Dumas, M., Barker, J. L., & Power, E. A. (2021). When does reputation lie? Dynamic feedbacks between costly signals, social capital and social prominence. *Philosophical Transactions of the Royal Society B*, 376(1838), 20200298.
- Dunbar, R. I. (2004). Gossip in evolutionary perspective. *Review of general psychology*, 8(2), 100-110.
- Dunfield, K., Kuhlmeier, V. A., O'Connell, L., & Kelley, E. (2011). Examining the diversity of prosocial behavior: Helping, sharing, and comforting in infancy. *Infancy*, 16(3), 227-247.
- Efferson, C., & Fehr, E. (2018). Simple moral code supports cooperation. *Nature*, 555, 169–170.
- Dunne, S., MacGabhann, L., Amering, M., & McGowan, P. (2018). “Making people aware and taking the stigma away”: Alleviating stigma and discrimination through dialogue. *Irish Journal of Applied Social Studies*, 18(1), 5.

- Efferson, C., McKay, R., & Fehr, E. (2020). The evolution of distorted beliefs vs. mistaken choices under asymmetric error costs. *Evolutionary Human Sciences*, 2, E27.
- Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: inter-and intranational differences. *Journal of personality and social psychology*, 81(5), 869.
- Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1), 49-98.
- Elsaadawy, N., & Carlson, E. N. (2022). Do you make a better or worse impression than you think?. *Journal of Personality and Social Psychology*, 123(6), 1407.
- Emler, N. (2019). *Human sociality and psychological foundations*. In Giardini, F., & Wittek, R. (Eds.), *The Oxford handbook of gossip and reputation* (pp. 47–68). Oxford University Press.
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological methods*, 12(2), 121.
- Engeler, N. C., & Raihani, N. J. (2024). Acting Without Considering Personal Costs Signals Trustworthiness in Helpers but Not Punishers. *Communications Psychology*, 2(1), 47.
- Engelmann, J. M., Herrmann, E., & Tomasello, M. (2012). Five-year olds, but not chimpanzees, attempt to manage their reputations. *PLoS One*, 7(10), e48433.
- Engelmann, J. M., Herrmann, E., & Tomasello, M. (2016). Preschoolers affect others' reputations through prosocial gossip. *British Journal of Developmental Psychology*, 34(3), 447–460.
- Engelmann, J. M., Over, H., Herrmann, E., & Tomasello, M. (2013). Young children care more about their reputation with ingroup members and potential reciprocators. *Developmental Science*, 16(6), 952-958.
- Engelmann, J. M., & Rapp, D. J. (2018). The influence of reputational concerns on children's prosociality. *Current Opinion in Psychology*, 20, 92-95.
- Engelmann, J. M., & Zeller, C. (2016). Doing the right thing for the wrong reason: reputation and moral behavior. In *The Routledge Handbook of Philosophy of the Social Mind* (pp. 263-277). Routledge.

- Engstrom, H. R., Laurin, K., Kay, N. R., & Human, L. J. (2023). Socioeconomic status and meta-perceptions: How markers of culture and rank predict beliefs about how others see us. *Personality and Social Psychology Bulletin*, 01461672231171435.
- Enke, B. (2019). Kinship, cooperation, and the evolution of moral systems. *The Quarterly Journal of Economics*, 134(2), 953–1019.
- Ensminger, J. (2000). 13. Experimental economics in the bush: why institutions matter. *Institutions, contracts, and organizations: Perspectives from new institutional economics*, 158.
- Epley, N., & Dunning, D. (2000). Feeling "holier than thou": are self-serving assessments produced by errors in self-or social prediction?. *Journal of personality and social psychology*, 79(6), 861.
- Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, 58(4), 723-733.
- Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations*, 19(2), 152-168.
- Eriksson, K., Andersson, P. A., & Strimling, P. (2017). When is it appropriate to reprimand a norm violation? The roles of anger, behavioral consequences, violation severity, and social distance. *Judgment and decision making*, 12(4), 396-407.
- Eriksson, K., Strimling, P., Andersson, P. A., Aveyard, M., Brauer, M., Gritskov, V., ... & Yamagishi, T. (2017). Cultural universals and cultural differences in meta-norms about peer punishment. *Management and Organization Review*, 13(4), 851-870.
- Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., ... & Van Lange, P. A. (2021). Perceptions of the appropriate response to norm violation in 57 societies. *Nature communications*, 12(1), 1481.
- Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5), 495-511.
- Espín, A. M., Branas-Garza, P., Herrmann, B., & Gamella, J. F. (2012). Patient and impatient punishers of free-riders. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749), 4923-4928.

- Evans, A. M., & Van De Calseyde, P. P. (2017). The effects of observed decision time on expectations of extremity and cooperation. *Journal of Experimental Social Psychology*, 68, 50-59.
- Everett, J. A. (2013). The 12 item social and economic conservatism scale (SECS). *PloS one*, 8(12), e82131.
- Everett, J. A., Faber, N. S., & Crockett, M. J. (2015). The influence of social preferences and reputational concerns on intergroup prosocial behaviour in gains and losses contexts. *Royal Society Open Science*, 2(12), 150546.
- Eysenck, H. J. (1976). *Sex and personality*. Austin: University of Texas Press.
- Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and individual differences*, 6(1), 21-29.
- Fabbri, M., & Carbonara, E. (2017). Social influence on third-party punishment: An experiment. *Journal of Economic Psychology*, 62, 204-230.
- Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, 73(6), 2017-2030.
- Fay, A. J., Jordan, A. H., & Ehrlinger, J. (2012). How social norms promote misleading social feedback and inaccurate self-assessment. *Social and Personality Psychology Compass*, 6(2), 206-216.
- Fehl, K., Sommerfeld, R. D., Semmann, D., Krambeck, H. J., & Milinski, M. (2012). I dare you to punish me—vendettas in games of cooperation. *PLoS One*, 7(9), e45093
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785–791.
- Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, 25(2), 63-87.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980-994.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137–140.
- Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928), 137.

- Feinberg, M., Willer, R., & Keltner, D. (2012). Flustered and faithful: Embarrassment as a signal of prosociality. *Journal of personality and social psychology*, 102(1), 81.
- Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological science*, 25(3), 656-664.
- FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature human behaviour*, 3(5), 426-435.
- Fincher, C. L., Thornhill, R., Murray, D. R., & Schaller, M. (2008). Pathogen prevalence predicts human cross-cultural variability in individualism/collectivism. *Proceedings of the Royal Society B: Biological Sciences*, 275(1640), 1279-1285.
- Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of personality and Social Psychology*, 38(6), 889.
- Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. *American Psychologist*, 48, 621–628.
- Fiske, S. T. (2014). Schema-triggered affect: Applications to social perception. In *Affect and cognition* (pp. 55-78). Psychology Press.
- Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6), 878–902.
- Fletcher, D., & Fletcher, D. (2018). *Why Model Averaging?* (pp. 1-29). Springer Berlin Heidelberg.
- Fombrun, C.J., & Rindova, V. (1996). Who's tops and who decides? The social construction of corporate reputations. *New York University, Stern School of Business, Working Paper*, 5–13.
- Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral ecology and sociobiology*, 65, 47-55.
- Francis, L. J. (1997). The relationship between Rosenberg's construct of self-esteem and Eysenck's two-dimensional model of personality. *Personality and individual differences*, 1(22), 139.

- Freckleton, R. P. (2011). Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology*, 65, 91-101.
- Freedman, G., & Flanagan, M. (2017). From dictators to avatars: Furthering social and personality psychology through game methods. *Social and personality psychology compass*, 11(12), e12368.
- Freeman, D., Garety, P. A., Bebbington, P. E., Smith, B., Rollinson, R., Fowler, D., ... & Dunn, G. (2005). Psychological investigation of the structure of paranoia in a non-clinical population. *The British Journal of Psychiatry*, 186(5), 427-435.
- Friedman, D., & Cassar, A. (2004). *Economists go to the laboratory: who, what, when, and why*. In *Economics Lab* (pp. 26-36). Routledge.
- Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E*, 78(2), 026117.
- Fu, G., Heyman, G. D., Qian, M., Guo, T., & Lee, K. (2016). Young children with a positive reputation to maintain are less likely to cheat. *Developmental Science*, 19(2), 275–283.
- Funder, D. C. (1995). On the accuracy of personality judgment: a realistic approach. *Psychological review*, 102(4), 652.
- Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception*. San Diego, CA: Academic Press.
- Funder, D. C. (2001). Accuracy in personality judgment: Research and theory concerning an obvious question. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace* (pp. 121–140).
- Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin*, 40(8), 986-997.
- Furr, R. M. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of personality*, 76(5), 1267-1316.
- Furr, R. M., & Funder, D. C. (2004). Situational similarity and behavioral consistency: Subjective, objective, variable-centered, and person-centered approaches. *Journal of Research in Personality*, 38(5), 421-447.

- Gächter, S., & Fehr, E. (1999). Collective action as a social exchange. *Journal of economic behavior & organization*, 39(4), 341-369.
- Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1518), 791–806.
- Galizzi, M. M., & Navarro-Martinez, D. (2019). On the external validity of social preference games: a systematic lab-field study. *Management Science*, 65(3), 976-1002.
- Gallrein, A. M. B., Weßels, N. M., Carlson, E. N., & Leising, D. (2016). I still cannot see it—A replication of blind spots in self-perception. *Journal of Research in Personality*, 60, 1-7.
- Garfield, Z. H., Hubbard, R. L., & Hagen, E. H. (2019). Evolutionary models of leadership. *Human Nature*, 30(1), 23-58.
- Garfield, Z. H., Schacht, R., Post, E. R., Ingram, D., Uehling, A., & Macfarlan, S. J. (2021). The content and structure of reputation domains across human societies: a view from the evolutionary social sciences. *Philosophical Transactions of the Royal Society B*, 376(1838), 20200296.
- Gehlbach, H., Brown, S. W., Ioannou, A., Boyer, M. A., Hudson, N., Niv-Solomon, A., ... & Janik, L. (2008). Increasing interest in social studies: Social perspective taking and self-efficacy in stimulating simulations. *Contemporary Educational Psychology*, 33(4), 894-914.
- Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., ... & Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033), 1100-1104.
- Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1), 33-48.
- Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56(2), 165-193.
- Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, 177, 122-141.



- Gervais, M. M. (2017). RICH economic games for networked relationships and communities: Development and preliminary validation in Yasawa, Fiji. *Field methods*, 29(2), 113-129.
- Giardini, F., & Wittek, R. (2019). *Gossip, Reputation and Sustainable Cooperation: Sociological Foundations*. Oxford University Press.
- Giardini, F., Balliet, D., Power, E. A., Számadó, S., & Takács, K. (2022). Four puzzles of reputation-based cooperation: Content, process, honesty, and structure. *Human Nature*, 33(1), 43-61.
- Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of personality and social psychology*, 78(2), 211.
- Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: biased assessments of others' ability to read one's emotional states. *Journal of personality and social psychology*, 75(2), 332.
- Gintis, H., Henrich, J., Bowles, S., Boyd, R., & Fehr, E. (2008). Strong Reciprocity and the Roots of Human Morality. *Social Justice Research*, 21(2), 241–253.
- Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of theoretical biology*, 213(1), 103-119.
- Goff, P. A., Jackson, M. C., Di Leone, B. A. L., Culotta, C. M., & DiTomasso, N. A. (2014). The essence of innocence: consequences of dehumanizing Black children. *Journal of personality and social psychology*, 106(4), 526.
- Goffman, E. (1956). Embarrassment and social organization. *American Journal of Sociology*, 62, 264-271.
- Goffman, E. (1958). *The presentation of self in everyday life*. University of Edinburgh Social Sciences Research Centre.
- Góis, A. R., Santos, F. P., Pacheco, J. M., & Santos, F. C. (2019). Reward and punishment in climate change dilemmas. *Scientific reports*, 9(1), 16193.
- Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, 24(1), 38-44.

Goodwin, G. P., & Benforado, A. (2015). Judging the goading ox: Retribution directed toward animals. *Cognitive science*, 39(3), 619-646.

Goodwin, S. A., Gubin, A., Fiske, S. T., & Yzerbyt, V. Y. (2000). Power can bias impression processes: Stereotyping subordinates by default and by design. *Group processes & intergroup relations*, 3(3), 227-256.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of personality and social psychology*, 106(1), 148.

Gordon, D. S., & Lea, S. E. (2016). Who punishes? The status of the punishers affects the perceived success of, and indirect benefits from, “moralistic” punishment. *Evolutionary Psychology*, 14(3), 1474704916658042.

Gordon, D. S., Madden, J. R., & Lea, S. E. (2014). Both loved and feared: third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PloS One*, 9(10), e110045.

Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, 84(11), 2809-2815.

Graham, M. E., & Cable, D. M. (2001). Consideration of the incomplete block design for policy-capturing research. *Organizational Research Methods*, 4(1), 26-45.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 83(6), 1420-1443.

Graziano, W. G., & Tobin, R. M. (2009). Agreeableness. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 46–61). The Guilford Press.

Grieco, D., Faillo, M., & Zarri, L. (2017). Enforcing cooperation in public goods games: Is one punisher enough?. *Journal of Economic Psychology*, 61, 55-73.

Gross, J., & De Dreu, C. K. (2019). The rise and fall of cooperation through reputation and group polarization. *Nature communications*, 10(1), 776.

Grossmann, I., Brienza, J. P., & Bobocel, D. R. (2017). Wise deliberation sustains cooperation. *Nature Human Behaviour*, 1(3), 0061.

- Grueber, C. E., Nakagawa, S., Laws, R. J., & Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *Journal of evolutionary biology*, 24(4), 699-711.
- Grueneisen, S., & Tomasello, M. (2017). Children coordinate in a recurrent social dilemma by taking turns and along dominance asymmetries. *Developmental psychology*, 53(2), 265.
- Grueneisen, S., Török, G., Wathiyage Don, A., & Ruggeri, A. (2024). Young children's adaptive partner choice in cooperation and competition contexts. *Child Development*, 95(3), 1023-1031.
- Grueneisen, S., & Warneken, F. (2022). The development of prosocial behaviour—from sympathy to strategy. *Current opinion in psychology*, 43, 323-328.
- Guala, F. (2008). Paradigmatic experiments: The ultimatum game from testing to measurement device. *Philosophy of Science*, 75(5), 658-669.
- Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and brain sciences*, 35(1), 1-15.
- Guala, F., & Mittone, L. (2010). Paradigmatic experiments: the dictator game. *The Journal of Socio-Economics*, 39(5), 578-584.
- Gurven, M., Allen-Arave, W., Hill, K., & Hurtado, M. (2000). “It's a wonderful life”: signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior*, 21(4), 263-282.
- Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical population biology*, 69(3), 339-348.
- Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8), 2395-2400.
- Hall, A. T., Blass, F. R., Ferris, G. R., & Massengale, R. (2004). Leader reputation and accountability in organizations: Implications for dysfunctional leader behavior. *The Leadership Quarterly*, 15(4), 515-536.
- Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450(7169), 557-559.

Harbaugh, W. T. (1998). What do donations buy?: A model of philanthropy based on prestige and warm glow. *Journal of Public Economics*, 67(2), 269–284.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, 316(5831), 1622–1625.

Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, 32(10), 1402–1413.

Hardy, R. A., & Norgaard, J. R. (2016). Reputation in the Internet black market: An empirical and theoretical analysis of the Deep Web. *Journal of Institutional Economics*, 12(3), 515–539.

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., ... & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794.

Hart, B. L., & Hart, L. A. (1992). Reciprocal allogrooming in impala, *Aepyceros melampus*. *Animal Behaviour*, 44(6), 1073–1083.

Haselton, M. G. (2007). Error Management Theory. In R. F. Baumeister and K. D. Vohs, editors, *Encyclopedia of social psychology*, 1, 311–312. Thousand Oaks, CA: Sage.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of personality and social psychology*, 78(1), 81–91.

Haselton, M. G., & Buss, D. M. (2003). Biases in social judgment: Design flaws or design features? In K. D. W. J. P. Forgas and W. von Hippel, *Responding to the social world: Implicit and explicit processes in social judgments and decisions*, 23–43. New York: Cambridge University Press.

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, 10(1), 47–66.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition*, 27(5), 733–763.

Hassin, R. R., Aarts, H., & Ferguson, M. J. (2005). Automatic goal inferences. *Journal of Experimental Social Psychology*, 41(2), 129–140.

- Hater, L., Elsaadawy, N., Biesanz, J. C., Breil, S. M., Human, L. J., Niemeyer, L. M., ... & Carlson, E. N. (2023). Examining individual differences in metaperceptive accuracy using the social meta-accuracy model. *Journal of Personality and Social Psychology*.
- Hauser, C., Perkmann, U., Puntischer, S., Walde, J., & Tappeiner, G. (2016). Trust works! Sources and effects of social capital in the workplace. *Social Indicators Research*, 128, 589-608.
- Heffner, J., & FeldmanHall, O. (2019). Why we don't always punish: Preferences for non-punitive responses to moral violations. *Scientific reports*, 9(1), 1-13.
- Henrich, J. (2017). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University
- Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of theoretical biology*, 208(1), 79-89.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and brain sciences*, 28(6), 795-815.
- Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., ... & Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *science*, 327(5972), 1480-1484.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and brain sciences*, 33(2-3), 111.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29-29.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, 33(2-3), 61-83.
- Henrich, J., & Henrich, N. (2014). Fairness without punishment: behavioral experiments in the Yasawa Island, Fiji. *Experimenting with social norms: Fairness and punishment in cross-cultural perspective*, 171-218.
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... & Ziker, J. (2006). Costly punishment across human societies. *Science*, 312(5781), 1767-1770.

- Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319(5868), 1362-1367.
- Herrmann, E., Engelmann, J. M., & Tomasello, M. (2019). Children engage in competitive altruism. *Journal of Experimental Child Psychology*, 179, 176-189.
- Herrmann, E., Keupp, S., Hare, B., Vaish, A., & Tomasello, M. (2013). Direct and indirect reputation formation in nonhuman great apes (*Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo pygmaeus*) and human children (*Homo sapiens*). *Journal of Comparative Psychology*, 127(1), 63–75.
- Heugens, P.P., Van Riel, C.B., & Van Den Bosch, F.A. (2004). Reputation management capabilities as decision rules. *Journal of Management Studies*, 41, 1349–1377.
- Heyman, G., Barner, D., Heumann, J., & Schenck, L. (2014). Children's sensitivity to ulterior motives when evaluating prosocial behavior. *Cognitive science*, 38(4), 683-700.
- Heyman, G. D., Fu, G., & Lee, K. (2007). Evaluating claims people make about themselves: The development of skepticism. *Child Development*, 78(2), 367–375.
- Hilbe, C., & Traulsen, A. (2012). Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Scientific reports*, 2(1), 1-4.
- Hilbe, C., Traulsen, A., Röhl, T., & Milinski, M. (2014). Democratic decisions establish stable authorities that overcome the paradox of second-order punishment. *Proceedings of the National Academy of Sciences*, 111(2), 752-756.
- Hirschmüller, S., Schmukle, S. C., Krause, S., Back, M. D., & Egloff, B. (2018). Accuracy of self-esteem judgments at zero acquaintance. *Journal of Personality*, 86(2), 308-319.
- Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, 148(3), 520.
- Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, 150(11), 2246.
- Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., ... & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring

preferences for intergroup inequality using the new SDO<sub>7</sub> scale. *Journal of personality and social psychology*, 109(6), 1003.

Hoffman, M., Hilbe, C., & Nowak, M. A. (2018). The signal-burying game can explain why we obscure positive traits and good deeds. *Nature Human Behaviour*, 2(6), 397-404.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7, 346–380.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences*, 112(6), 1727–1732.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin*, 44(12), 1697-1711.

Hogan, M. J., Chiricos, T., & Gertz, M. (2005). Economic insecurity, blame, and punitive attitudes. *Justice Quarterly*, 22(3), 392-412.

Holoien, D. S., & Fiske, S. T. (2013). Downplaying positive impressions: Compensation between warmth and competence in impression management. *Journal of Experimental Social Psychology*, 49(1), 33-41.

Hooper, P. L., Kaplan, H. S., & Boone, J. L. (2010). A theory of leadership in human cooperative groups. *Journal of Theoretical Biology*, 265(4), 633-646.

Horita, Y. (2010). Punishers may be chosen as providers but not as recipients. *Letters on Evolutionary Behavioral Science*, 1(1), 6-9.

Horton, L. E., Barrantes-Vidal, N., Silvia, P. J., & Kwapil, T. R. (2014). Worries about being judged versus being harmed: Disentangling the association of social anxiety and paranoia with schizotypy. *PloS one*, 9(6), e96269.

Houser, D., Schunk, D., & Winter, J. (2010). Distinguishing trust from risk: An anatomy of the investment game. *Journal of economic behavior & organization*, 74(1-2), 72-81.

Hox, J. J., Kreft, I. G., & Hermkens, P. L. (1991). The analysis of factorial surveys. *Sociological Methods & Research*, 19(4), 493-510.

Hughes, R., & Huby, M. (2002). The application of vignettes in social and nursing research. *Journal of advanced nursing*, 37(4), 382-386.

Irwin, K., & Horne, C. (2013). A normative explanation of antisocial punishment. *Social Science Research*, 42(2), 562–570.

Jaeger, B. (2017). [r2glmm: Computes R Squared for Mixed \(Multilevel\) Models.](https://CRAN.R-project.org/package=r2glmm.57364)  
[https://CRAN.R-project.org/package=r2glmm.57364.](https://CRAN.R-project.org/package=r2glmm.57364)

Janoff-Bulman, R., Sheikh, S., & Baldacci, K. G. (2008). Mapping moral motives: Approach, avoidance, and political orientation. *Journal of Experimental Social Psychology*, 44(4), 1091-1099.

Jeffreys, H. (1939). *Theory of Probability*. Oxford: Clarendon Press.

Jia, Z. T., & McMahon, M. J. (2020). Being watched in an investment game setting: Behavioral changes when making risky decisions. *Journal of Behavioral and Experimental Economics*, 88, 101593.

Jin, S., Balliet, D., Romano, A., Spadaro, G., Van Lissa, C. J., Agostini, M., ... & Leander, N. P. (2021). Intergenerational conflicts of interest and prosocial behavior during the COVID-19 pandemic. *Personality and Individual Differences*, 171, 110535.

John, O. P., & Srivastava, S. (1999). *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives* (Vol. 2, pp. 102-138). Berkeley: University of California.

Johnson, D. (2009). Anger about crime and support for punitive criminal justice policies. *Punishment & Society*, 11(1), 51-66.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, 102(3), 192–194.

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in ecology & evolution*, 19(2), 101-108.

Johnson, S. G., & Park, S. Y. (2021). Moral signaling through donations of money and time. *Organizational Behavior and Human Decision Processes*, 165, 183-196.

Jolly, E., & Chang, L. J. (2021). Gossip drives vicarious learning and facilitates social connection. *Current Biology*, 31(12), 2539-2549.

Jones, D., & Linardi, S. (2012). *Gender differences in reputation concerns: Prosocial behavior in the field & lab*. Working paper, Tech rep.



- Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. (2016b). Third-party punishment as a costly signal of trustworthiness. *Nature*, 530(7591), 473-476.
- Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016a). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, 113(31), 8658-8663.
- Jordan, J., & Kteily, N. (2020). Reputation fuels moralistic punishment that people judge to be questionably merited. Preprint.
- Jordan, J. J., & Kteily, N. S. (2023). How reputation does (and does not) drive people to punish without looking. *Proceedings of the National Academy of Sciences*, 120(28), e2302475120.
- Jordan, J. J., & Rand, D. G. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of theoretical biology*, 421, 189-202.
- Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of personality and social psychology*, 118(1), 57.
- Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why Do I Hate Hypocrites? Evidence for a Theory of False Signaling. *Psychological science*, 28(3), 356-368.
- Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, 89, 899-913.
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond*. Routledge.
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5), 1449-1475.
- Kalogiannidis, S. (2020). Impact of effective business communication on employee performance. *European Journal of Business and Management Research*, 5(6).
- Kanakogi, Y., Miyazaki, M., Takahashi, H., Yamamoto, H., Kobayashi, T., & Hiraki, K. (2022). Third-party punishment by preverbal infants. *Nature human behaviour*, 6(9), 1234-1242.

- Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies*, 59(1), 63-80.
- Kaplan, S. A., Santuzzi, A. M., & Ruscher, J. B. (2009). Elaborative metaperceptions in outcome-dependent situations: The diluted relationship between default self-perceptions and metaperceptions. *Social Cognition*, 27(4), 601-614.
- Kawamura, Y., & Kusumi, T. (2020). Altruism does not always lead to a good reputation: A normative explanation. *Journal of Experimental Social Psychology*, 90, 104021.
- Kawash, G. F. (1982). A structural analysis of self-esteem from pre-adolescence through young adulthood: Anxiety and extraversion as agents in the development of self-esteem. *Journal of Clinical Psychology*, 38(2), 301-311.
- Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska symposium on motivation*. University of Nebraska Press.
- Kelsey, C., Grossmann, T., & Vaish, A. (2018). Early reputation management: Three-year-old children are more generous following exposure to eyes. *Frontiers in Psychology*, 9, 345016.
- Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological review*, 110(2), 265.
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York, NY: Guilford Press.
- Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and social psychology review*, 8(3), 265-280.
- Kenny, D. A., & DePaulo, B. M. (1993). Do people know how others view them? An empirical and theoretical account. *Psychological Bulletin*, 114(1), 145–161.
- Kenny, D. A., & La Voie, L. (1984). The social relations model. *Advances in Experimental Social Psychology*, 18, 141–182.
- Kervyn, N., Fiske, S., & Yzerbyt, V. (2015). Forecasting the primary dimension of social perception. *Social Psychology*.
- Kim, B., Park, E., Cameron, G. T., MEADOWS, C., OTT, H., & XIAO, A. (2017). Transparent communication efforts inspire confident, even greater, employee performance. *Asian Journal of Public Relations*, 1(1), 9-31.

- King, A., & Maruna, S. (2009). Is a conservative just a liberal who has been mugged? Exploring the origins of punitive views. *Punishment & Society*, 11(2), 147-169.
- Kirgios, E. L., Chang, E. H., Levine, E. E., Milkman, K. L., & Kessler, J. B. (2020). Forgoing earned incentives to signal pure motives. *Proceedings of the National Academy of Sciences*, 117(29), 16891-16897.
- Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: free riding may be thwarted by second-order reward rather than by punishment. *Journal of personality and social psychology*, 95(4), 826.
- Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: confusion or a heuristic?. *Evolution and human behavior*, 21(6), 411-427.
- Klein, S. A., & Hilbig, B. E. (2019). On the lack of real consequences in consumer choice research—and its consequences. *Experimental Psychology*, 66(1), 68.
- Kleinlogel, E. P., Renier, L., Mast, M. S., & Toma, C. (2020). I think that I made a good impression!. *Social Psychology*, 51(6), 2151-2590.
- Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual review of sociology*, 24(1), 183-214.
- Konrath, S., Meier, B. P., & Bushman, B. J. (2014). Development and validation of the single item narcissism scale (SINS). *PLOS one*, 9(8), e103469.
- Köymen, B., Lieven, E., Engemann, D. A., Rakoczy, H., Warneken, F., & Tomasello, M. (2014). Children's norm enforcement in their interactions with peers. *Child development*, 85(3), 1108-1122.
- Kraft-Todd, G., Kleiman-Weiner, M., & Young, L. (2020). Differential discounting of virtue signaling: public virtue is perceived less favorably than private virtue for generosity but not impartiality. *PsyArXiv*.
- Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are punishment and reputation for? *PLoS ONE*, 7(9), e45662.
- Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological science*, 27(3), 405-418.

- Kriss, P. H., Weber, R. A., & Xiao, E. (2016). Turning a blind eye, but not the other cheek: On the robustness of costly punishment. *Journal of Economic Behavior & Organization*, 128, 159-177.
- Krolak-Schwerdt, S., Hörstermann, T., Glock, S., & Böhmer, I. (2018). Teachers' assessments of students' achievements: The ecological validity of studies using case vignettes. *The Journal of Experimental Education*, 86(4), 515-529.
- Kruger, J., & Gilovich, T. (1999). " Naive cynicism" in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology*, 76(5), 743.
- Kupfer, T. R., & Giner-Sorolla, R. (2017). Communicating moral motives: The social signaling function of disgust. *Social Psychological and Personality Science*, 8(6), 632-640.
- Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, 28(2), 75–84.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82(1), 1-26.
- Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, 76(2), 225-237.
- Lapsley, D. K., & Lasky, B. (2001). Prototypic moral character. *Identity: An International Journal of Theory and Research*, 1(4), 345-363.
- Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 151, 61-72.
- Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, 93(2), 234.
- Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annu. Rev. Psychol.*, 58, 317-344.
- Lee, R. B. (1979). *The! Kung San: Men, women and work in a foraging society*. Cambridge University Press.

- Lee, S. Y., & Jang, K. (2019). Antecedents of impression management motivations on social network sites and their link to social anxiety. *Media Psychology*, 22(6), 890-904.
- Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.
- Lee, Y. E., & Warneken, F. (2020). Children's evaluations of third-party responses to unfairness: Children prefer helping over punishment. *Cognition*, 205, 104374.
- Lees, J., Young, L., & Waytz, A. (2022). Morally questionable actors' meta-perceptions are accurate but overly positive. *Journal of Experimental Social Psychology*, 102, 104371.
- Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London, B: Biological Sciences*, 268(1468), 745–753.
- Leimgruber, K. L., Shaw, A., Santos, L. R., & Olson, K. R. (2012). Young children are more generous when others are aware of their actions. *PloS one*, 7(10), e48292.
- [Lenth, R. \(2019\). emmeans: Estimated Marginal Means, aka Least-Squares Means. https://CRAN.R-project.org/package=emmeans.](https://CRAN.R-project.org/package=emmeans)
- Lenth, R. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.5. <https://CRAN.R-project.org/package=emmeans>.
- Lenth, R. (2023). emmeans: Estimated Marginal Means, aka Least-Squares Means.(v1. 8.4–1) Available at: <https://CRAN.R-project.org/package=emmeans>.
- Lesorogol, C. (2005). Experiments and ethnography: combining methods for better understanding of behavior and change. *Current Anthropology*, 46(1), 129-136.
- Letzring, T. D., & Funder, D. C. (2018). Interpersonal accuracy in trait judgments. *The SAGE handbook of personality and individual differences*, 3, 253-282.
- Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126, 88-106.
- Levy, N. (2021). Virtue signalling is virtuous. *Synthese*, 198(10), 9545-9562.
- Li, Z., Hu, G., Xu, L., & Li, Q. (2021). Third-party punishment or compensation? It depends on the reputational benefits. *Frontiers in Psychology*, 12, 676064.

- Li, Y., Luo, J., Niu, H., & Ye, H. (2023). When punishers might be loved: fourth-party choices and third-party punishment in a delegation game. *Theory and Decision*, 94(3), 423-465.
- Li, Y., & Mifune, N. (2022). Non-cooperation motivation is irrelevant: Negative evaluations of punishment in the public goods game, positive evaluations in the third-party punishment game. Preprint.
- Lightner, A. D., Barclay, P., & Hagen, E. H. (2017). Radical framing effects in the ultimatum game: the impact of explicit culturally transmitted frames on economic decision-making. *Royal Society open science*, 4(12), 170543.
- Lin, M. H., Kwan, V. S., Cheung, A., & Fiske, S. T. (2005). Stereotype content model explains prejudice for an envied outgroup: Scale of anti-Asian American stereotypes. *Personality and Social Psychology Bulletin*, 31(1), 34-47.
- Little, R. J., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* . Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons], third ed.
- Liu, Q., Jiang, H., Pan, Z., Han, Q., Peng, Z., & Li, Q. (2024, March). BiasEye: A Bias-Aware Real-time Interactive Material Screening System for Impartial Candidate Assessment. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (pp. 325-343).
- Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366), 1038-1041.
- Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, 109(453), 25-34.
- Loke, I. C., Heyman, G. D., Forgie, J., McCarthy, A., & Lee, K. (2011). Children's moral evaluations of reporting the transgressions of peers: age differences in evaluations of tattling. *Developmental Psychology*, 47, 1757–1762.
- Loke, I., Heyman, G. D., Itakura, S., Toriyama, R., & Lee, K. (2014). Japanese and American children's moral evaluations of reporting on transgressions. *Developmental Psychology*, 50, 1520–1531.
- Lotem, A., Fishman, M. A., & Stone, L. (2003). From reciprocity to unconditional altruism through signalling benefits. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1511), 199-205.

- Lüdtke, D. (2021). SjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.10. <https://CRAN.R-project.org/package=sjPlot>.
- Ludwick, R., & Zeller, R. A. (2001). The factorial survey: An experimental method to replicate real world problems. *Nursing research*, 50(2), 129-133.
- Lukacs, P. M., Burnham, K. P., & Anderson, D. R. (2010). Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*, 62, 117-125.
- Luong, R., & Lomanowska, A. M. (2022). Evaluating Reddit as a crowdsourcing platform for psychology research projects. *Teaching of Psychology*, 49(4), 329-337.
- Lyle, H. F., & Smith, E. A. (2014). The reputational and social network benefits of prosociality in an Andean community. *Proceedings of the National Academy of Sciences*, 111(13), 4820–4825.
- Lyon, B. E., & Montgomerie, R. (2012). Sexual selection is a form of social selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1600), 2266-2273.
- Macfarlan, S. J., & Lyle, H. F. (2015). Multiple reputation domains and cooperative behaviour in two Latin American communities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1683), 20150009.
- Macfarlan, S. J., Remiker, M., & Quinlan, R. (2012). Competitive altruism explains labor exchange variation in a Dominican community. *Current Anthropology*, 53(1), 118–124.
- Macfarlan, S. J., Quinlan, R., & Remiker, M. (2013). Cooperative behaviour and prosocial reputation dynamics in a Dominican village. *Proceedings of the Royal Society B: Biological Sciences*, 280(1761), 20130557.
- Maeda, K., Kumai, Y., & Hashimoto, H. (2022). Potential influence of decision time on punishment behavior and its evaluation. *Frontiers in Psychology*, 4597.
- Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, 102(4), 661.
- Maner, J. K., & Mead, N. L. (2010). The essential tension between leadership and power: when leaders sacrifice group goals for the sake of self-interest. *Journal of personality and social psychology*, 99(3), 482.

Manrique, H. M., Walker, M. J. (2017). *Early evolution of human memory*. Cham, Switzerland: Palgrave Macmillan.

Manrique, H. M., Zeidler, H., Roberts, G., Barclay, P., Walker, M., Samu, F., ... & Raihani, N. (2021). The psychological foundations of reputation-based cooperation. *Philosophical Transactions of the Royal Society B*, 376(1838), 20200287

Marlowe, F. W. (2009). Hadza cooperation: second-party punishment, yes; third-party punishment, no. *Human Nature*, 20, 417-430.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences*, 275(1634), 587–592.

Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M., & Tracer, D. (2011). The 'spiteful' origins of human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 278(1715), 2159–2164.

Marshall, J., Gollwitzer, A., & Bloom, P. (2022). Why do children and adults think other people punish?. In press at *Developmental psychology*.

Marshall, J., & McAuliffe, K. (2022). Children as assessors and agents of third-party punishment. *Nature Reviews Psychology*, 1(6), 334-344.

Marshall, J., Mermin-Bunnell, K., & Bloom, P. (2020). Developing judgments about peers' obligation to intervene. *Cognition*, 201, 104215.

Marshall, J., Yudkin, D. A., & Crockett, M. J. (2021). Children punish third parties to satisfy both consequentialist and retributive motives. *Nature Human Behaviour*, 5(3), 361-368.

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, 147, 133-143.

Martin, J. W., Jordan, J. J., Rand, D. G., & Cushman, F. (2019). When do we punish people who don't?. *Cognition*, 193, 104040.

Mastroianni, A. M., Cooney, G., Boothby, E. J., & Reece, A. G. (2021). The liking gap in groups and teams. *Organizational Behavior and Human Decision Processes*, 162, 109-122.



- Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences*, 108(28), 11375-11380.
- Mazzocco, P. J., Alicke, M. D., & Davis, T. L. (2004). On the robustness of outcome bias: Noconstraint by prior culpability. *Basic and Applied Social Psychology*, 26(2-3), 131-146.
- McAuliffe, K., Blake, P. R., & Warneken, F. (2020). Costly fairness in children is influenced by who is watching. *Developmental psychology*, 56(4), 773.
- McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, 134, 1-10.
- McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, 36(1), 1-15.
- McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior*, 31(5), 309–319.
- McKenzie, C. R., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, 54(1), 33-61.
- McNamara, J. M., & Barta, Z. (2020). Behavioural flexibility and reputation formation. *Proceedings of the Royal Society B: Biological Sciences*, 287(1939), 20201758.
- McNamara, J. M., Barta, Z., Fromhage, L., & Houston, A. I. (2008). The coevolution of choosiness and cooperation. *Nature*, 451(7175), 189–192.
- Mehl, M. R., & Conner, T. S. (2011). *Handbook of research methods for studying daily life*. Guilford Press.
- Melis, A. P., & Raihani, N. J. (2023). The cognitive challenges of cooperation in human and nonhuman animals. *Nature Reviews Psychology*, 2(9), 523-536.
- Melis, A. P., & Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1553), 2663–2674.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Mifune, N., Li, Y., & Okuda, N. (2020). The evaluation of second-and third-party punishers. *Letters on Evolutionary Behavioral Science*, 11(1), 6-9.

- Milinski, M., Semmann, D., Bakker, T. C., & Krambeck, H. J. (2001). Cooperation through indirect reciprocity: image scoring or standing strategy?. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1484), 2495-2501.
- Milinski, M., Semmann, D., & Krambeck, H. (2002). Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1494), 881-883.
- Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the ‘tragedy of the commons’. *Nature*, 415(6870), 424-426.
- Milinski, M., Semmann, D., Krambeck, H. J., & Marotzke, J. (2006). Stabilizing the Earth’s climate is not a losing game: Supporting evidence from public goods experiments. *Proceedings of the National Academy of Sciences*, 103(11), 3994-3998.
- Milinski, M., & Wedekind, C. (1998). Working memory constrains human cooperation in the Prisoner’s Dilemma. *Proceedings of the National Academy of Sciences*, 95(23), 13755-13758.
- Miller, D. T. (1999). The norm of self-interest. *American Psychologist*, 54(12), 1053.
- Miller, D. T., & Ratner, R. K. (1998). The disparity between the actual and assumed power of self-interest. *Journal of personality and social psychology*, 74(1), 53.
- Miller, W. B. (1973). Ideology and criminal justice policy: Some current issues. *J. Crim. L. & Criminology*, 64, 141.
- Mills, C. M., & Keil, F. C. (2005). The development of cynicism. *Psychological Science*, 16(5), 385–390.
- Molenmaker, W. E., de Kwaadsteniet, E. W., & van Dijk, E. (2016). The impact of personal responsibility on the (un) willingness to punish non-cooperation and reward cooperation. *Organizational Behavior and Human Decision Processes*, 134, 1-15.
- Molho, C., & Wu, J. (2021). Direct punishment and indirect reputation-based tactics to intervene against offences. *Philosophical Transactions of the Royal Society B*, 376(1838), 20200289.
- Molho, C., Tybur, J. M., Van Lange, P. A., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature communications*, 11(1), 1-9.

- Molnar, A., Chaudhry, S., & Loewenstein, G. F. (2020). 'It's not about the money. It's about sending a message!': Unpacking the components of revenge. *SSRN Electronic Journal*. Retrieved from <https://www.ssrn.com/abstract,3524910>.
- Mooijman, M., & Graham, J. (2018). Unjust punishment in organizations. *Research in Organizational Behavior*, 38, 95-106.
- Moore-Berg, S. L., Ankori-Karlinsky, L. O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences*, 117(26), 14864-14872.
- Morey, R., & Rouder, J. (2022). BayesFactor: Computation of Bayes Factors for Common Designs (Version 0.9.12-4.4) [Software]. Retrieved from <https://CRAN.R-project.org/package=BayesFactor>
- Moskowitz, G. B., & Olcaysoy Okten, I. (2016). Spontaneous goal inference (SGI). *Social and Personality Psychology Compass*, 10(1), 64-80.
- Munniksma, A., Flache, A., Verkuyten, M., & Veenstra, R. (2012). Parental acceptance of children's intimate ethnic outgroup relations: The role of culture, status, and family reputation. *International Journal of Intercultural Relations*, 36(4), 575-585.
- Murnighan, J. K., & Wang, L. (2016). The social world as an experimental game. *Organizational Behavior and Human Decision Processes*, 136, 80-94.
- Nakamaru, M., & Iwasa, Y. (2006). The coevolution of altruism and punishment: role of the selfish punisher. *Journal of Theoretical Biology*, 240(3), 475-488.
- Nax, H. H., Perc, M., Szolnoki, A., & Helbing, D. (2015). Stability of cooperation under image scoring in group interactions. *Scientific reports*, 5(1), 12145.
- Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, 29(4), 242-248.
- Nesse, R. M. (2007). Runaway social selection for displays of partner value and altruism. *Biological Theory*, 2, 143-155.
- Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological science*, 25(3), 648-655.

- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301
- Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, 92(1), 91–112.
- Nikiforakis, N., & Engelmann, D. (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior & Organization*, 78(3), 319-332.
- Nikiforakis, N., & Normann, H. T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, 11, 358-369.
- Nishi, A., Christakis, N. A., Evans, A. M., O'Malley, A. J., & Rand, D. G. (2016). Social environment shapes the speed of cooperation. *Scientific reports*, 6(1), 1-10.
- Noë, R., & Hammerstein, P. (1994). Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, 35(1), 1–11.
- Noë, R., & Hammerstein, P. (1995). Biological markets. *Trends in Ecology & Evolution*, 10(8), 336–339.
- Nosenzo, D., & Sefton, M. (2014). *Promoting Cooperation: The Distribution of Reward and Punishment Power*. In *Social Dilemmas: New Perspectives on Reward and Punishment*. Oxford: Oxford University Press, p. 32.
- Nowak, M.A., & Sigmund, K. (1998a). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685), 573–577.
- Nowak, M. A., & Sigmund, K. (1998b). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, 194(4), 561–574.
- Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063), 1291-1298.
- Oberg, A. L., & Mahoney, D. W. (2007). *Linear Mixed Effects Models*. In: Ambrosius, W.T. (eds) *Topics in Biostatistics. Methods in Molecular Biology™*, vol 404. Humana Press.
- O'Connor, K., Effron, D. A., & Lucas, B. J. (2020). Moral cleansing as hypocrisy: When private acts of charity make you feel better than you deserve. *Journal of personality and social psychology*, 119(3), 540.

- Odine, M. (2015). Communication Problems in Management. *Journal of Emerging Issues in Economics, Finance & Banking*, 4(2), 1615-1630.
- O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 276(1655), 323-329.
- Ohtsuki, H., & Iwasa, Y. (2004). How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of theoretical biology*, 231(1), 107-120.
- Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, 239(4), 435–444.
- Ohtsuki, H., & Iwasa, Y. (2007). Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *Journal of theoretical biology*, 244(3), 518-531.
- Ohtsuki, H., Iwasa, Y., & Nowak, M. A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature*, 457(7225), 79-82.
- Olivola, C. Y., & Shafir, E. (2013). The martyrdom effect: When pain and effort increase prosocial contributions. *Journal of behavioral decision making*, 26(1), 91-105.
- Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, 10(8), 531-539.
- Osgood, J. M. (2017). Is revenge about retributive justice, deterring harm, or both?. *Social and Personality Psychology Compass*, 11(1), e12296.
- Over, H. (2016). The origins of belonging: Social motivation in infants and young children. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1686), 20150072.
- Ozono, H., Kamijo, Y., & Shimizu, K. (2020). The role of peer reward and punishment for public goods problems in a localized society. *Scientific reports*, 10(1), 8211.
- Ozono, H., & Watabe, M. (2012). Reputational benefit of punishment: comparison among the punisher, rewarder, and non-sanctioner. *Letters on Evolutionary Behavioral Science*, 3(2), 21-24.

- Pacheco, J. M., Santos, F. C., & Chalub, F. A. C. (2006). Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS computational biology*, 2(12), e178.
- Pagliaro, S., Cavazza, N., Guidetti, M., Alparone, F. R., & Giannella, V. A. (2016). Concern for personal reputation within groups: the effect of accountability and fear of social exclusion. *Ricerche di Psicologia*, 39(1), 9-19.
- Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of theoretical biology*, 224(1), 115-126.
- Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, 432(7016), 499-502.
- Paunonen, S. V., & Hong, R. Y. (2015). On the properties of personality traits. In M. Mikulincer, P. R. Shaver, M. L. Cooper, & R. J. Larsen (Eds.), *APA handbook of personality and social psychology, Vol. 4. Personality processes and individual differences* (pp. 233–259).
- Pedersen, E. J., McAuliffe, W. H., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, 147(4), 514.
- Pedersen, E. J., McAuliffe, W. H., Shah, Y., Tanaka, H., Ohtsubo, Y., & McCullough, M. E. (2020). When and why do third parties punish outside of the lab? *A cross-cultural recall study. Social Psychological and Personality Science*, 11(6), 846-853
- Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences*, 280(1758), 20122723.
- Peng, J., & Yang, Y. (2022). On improvability of model selection by model averaging. *Journal of Econometrics*, 229(2), 246-262.
- Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: Multilevel perspectives. *Annual Review of Psychology*, 56, 365-392.
- Pennisi, E. (2009). On the origin of cooperation. *Science*, 325, 1196-1199.
- Peters, L., Sunderland, M., Andrews, G., Rapee, R. M., & Mattick, R. P. (2012). Development of a short form Social Interaction Anxiety (SIAS) and Social Phobia Scale (SPS) using

nonparametric item response theory: the SIAS-6 and the SPS-6. *Psychological assessment*, 24(1), 66.

Pfattheicher, S. (2015). A regulatory focus perspective on reputational concerns: The impact of prevention-focused self-regulation. *Motivation and Emotion*, 39(6), 932-942.

Pfattheicher, S., Böhm, R., & Kesberg, R. (2018). The advantage of democratic peer punishment in sustaining cooperation within groups. *Journal of Behavioral Decision Making*, 31(4), 562-571.

Piazza, J., & Bering, J. (2010). The coevolution of secrecy and stigmatization: evidence from the content of distressing secrets. *Human Nature*, 21(3), 290-308.

Piazza, J., & Bering, J. M. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, 6(3), 147470490800600314.

[PIEL Survey. https://pielsurvey.org/](https://pielsurvey.org/)

Pinker, S. (2022). *Rationality: What it is, why it seems scarce, why it matters*. Penguin.

Pisor, A. C., Fessler, D. M. (2012). Importing social preferences across contexts and the pitfall of over-generalization across theories. Commentary on Guala ‘Weak or strong? What punishment experiments do (and do not) demonstrate’. *Behavioral and Brain Sciences*, 35(1), 34.

Pisor, A. C., Gervais, M. M., Purzycki, B. G., & Ross, C. T. (2020). Preferences and constraints: the value of economic games for studying human behaviour. *Royal Society open science*, 7(6), 192090.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, 14(3), 267–272.

Pope, D. G., Price, J., & Wolfers, J. (2018). Awareness reduces racial bias. *Management Science*, 64(11), 4988-4995.

Posner, R. (1997). Social norms and the law: An economic approach. *American Economic Review*, 87, 365-369.

Power, E. A., & Ready, E. (2018). Building bigness: Reputation, prominence, and social capital in rural South India. *American Anthropologist*, 120(3), 444–459.

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, 23(3), 203-231.

Price, M. E., & Van Vugt, M. (2014). The evolution of leader–follower reciprocity: the theory of service-for-prestige. *Frontiers in human neuroscience*, 8, 363.

Pruitt, D. G., & Kimmel, M. J. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual review of psychology*, 28(1), 363-392.

Przepiorka, W., & Liebe, U. (2016). Generosity is a sign of trustworthiness—the punishment of selfishness is not. *Evolution and human behavior*, 37(4), 255-262.

Przepiorka, W., Norbutas, L., & Corten, R. (2017). Order without law: Reputation promotes cooperation in a cryptomarket for illegal drugs. *European Sociological Review*, 33(6), 752–764.

Quillien, T. (2019). Universal modesty in signal-burying games. *Proceedings of the Royal Society B: Biological Sciences*, 286(1906), 20190985.

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge university press.

[R Core Team \(2019\). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.56761.](https://www.R-project.org/.56761)

R Core Team (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/.56761>.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>.

Radkani, S., Tenenbaum, J., & Saxe, R. (2022). Modeling punishment as a rational communicative social action. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44, No. 44).

Rai, T. S. (2022). Material benefits crowd out moralistic punishment. *Psychological Science*, 33(5), 789-797.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, 118(1), 57.



- Raihani, N. J., & Bell, V. (2018). Conflict and cooperation in paranoia: a large-scale behavioural experiment. *Psychological Medicine*, 48(9), 1523-1531.
- Raihani, N. J., & Bshary, R. (2015a). Why humans might help strangers. *Frontiers in behavioral neuroscience*, 9, 39.
- Raihani, N. J., & Bshary, R. (2015b). The reputation of punishers. *Trends in Ecology & Evolution*, 30(2), 98–103.
- Raihani, N. J., & Bshary, R. (2015c). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution*, 69(4), 993-1003.
- Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences*, 1.
- Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science*, 327(5962), 171-171.
- Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, 8(5), 802-804.
- Raihani, N. J., & Power, E. A. (2021). No good deed goes unpunished: the social costs of prosocial behaviour. *Evolutionary Human Sciences*, 3, e40.
- Raihani, N. J., & Smith, S. (2015). Competitive helping in online giving. *Current Biology*, 25(9), 1183–1186.
- Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in ecology & evolution*, 27(5), 288-295.
- Rammstedt, B. (2007). The 10-item big five inventory. *European Journal of Psychological Assessment*, 23(3), 193-201.
- Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality*, 41(1), 203-212.
- Rand, D. G., Armao IV, J. J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology*, 265(4), 624-632.

- Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in cognitive sciences*, 17(8), 413-425.
- Rand, D. G., Ohtsuki, H., & Nowak, M. A. (2009). Direct reciprocity with costly punishment: Generous tit-for-tat prevails. *Journal of theoretical biology*, 256(1), 45-57.
- Ratner, R. K., & Miller, D. T. (2001). The norm of self-interest and its effects on social action. *Journal of personality and social psychology*, 81(1), 5.
- Ready, R. E., Clark, L. A., Watson, D., & Westerhouse, K. (2000). Self-and peer-reported personality: Agreement, trait ratability, and the “self-based heuristic”. *Journal of research in personality*, 34(2), 208-224.
- Reb, J., Goldman, B. M., Kray, L. J., & Cropanzano, R. (2006). Different wrongs, different remedies? Reactions to organizational remedies after procedural and interactional injustice. *Personnel Psychology*, 59(1), 31-64.
- Redhead, D., Dhaliwal, N., & Cheng, J. T. (2021). Taking charge and stepping in: Individuals who punish are rewarded with prestige and dominance. *Social and Personality Psychology Compass*, 15(2), e12581.
- Reeder, G. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological inquiry*, 20(1), 1-18.
- Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, 86(1), 61.
- Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics*, 9(2), 79–101.
- Rhee, M., & Haunschild, P.R. (2006). The liability of good reputation: a study of product recalls in the US automobile industry. *Organization Science*, 17, 101–117.
- Richards, S. A. (2005). Testing ecological theory using the information-theoretic approach: examples and cautionary results. *Ecology*, 86(10), 2805-2814.
- Richards, S. A., Whittingham, M. J., & Stephens, P. A. (2011). Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behavioral Ecology and Sociobiology*, 65, 77-89.

- Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394), 427-431.
- Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, 70(4), 901–908.
- Roberts, G. (2020). Honest signaling of cooperative intentions. *Behavioral Ecology*, 31(4), 922-932.
- Roberts, G., Raihani, N., Bshary, R., Manrique, H. M., Farina, A., Samu, F., & Barclay, P. (2021). The benefits of being seen to help others: indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society B*, 376(1838), 20200290
- Robinett, T. L., & Malhotra, S. (2019). Conservative worldview and its influence on punishment and bioethical public health issues. *Ethics, Medicine and Public Health*, 8, 120-126.
- Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, 444(7120), 718-723.
- Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the National Academy of Sciences*, 108(45), 18307-18312.
- Rogers, T., Goldstein, N.J., & Fox, C.R. (2018). Social mobilization. *Annual Review of Psychology*, 69, 357–381.
- Romano, A., Giardini, F., Columbus, S., de Kwaadsteniet, E. W., Kisfalusi, D., Triki, Z., ... & Hagel, K. (2021). Reputation and socio-ecology in humans. *Philosophical Transactions of the Royal Society B*, 376(1838), 20200295.
- Rossano, F., Rakoczy, H., & Tomasello, M. (2011). Young children's understanding of violations of property rights. *Cognition*, 121(2), 219-227.
- Rowling, J. K. (2014). *Harry potter and the goblet of fire*. Bloomsbury Childrens Books.
- Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, 5(1), 2-14.
- Rozin, P. (2006). Domain denigration and process preference in academic psychology. *Perspectives on Psychological Science*, 1(4), 365-376.

- Ruttan, L. M., & Borgerhoff Mulder, M. (1999). Are East African pastoralists truly conservationists?. *Current anthropology*, 40(5), 621-652.
- Sabini, J., Garvey, B., & Hall, A. L. (2001). Shame and Embarrassment Revisited. *Personality and Social Psychology Bulletin*, 27(1), 104-117.
- Sadia, A., Salleh, B. M., Kadir, Z. A., & Sanif, S. (2016). The relationship between organizational communication and employees productivity with new dimensions of effective communication flow. *Journal of business and social review in emerging economies*, 2(2), 93-100.
- Salcedo García, J. C. (2020). Moralistic punishment signaling as a function of proportionality. [Doctoral Dissertation, Universidad de los Andes] <http://hdl.handle.net/1992/48393>
- Salcedo, J. C., & Jimenez-Leal, W. (2024). Severity and deservedness determine signalled trustworthiness in third party punishment. *British Journal of Social Psychology*, 63(1), 453-471.
- Samu, F., Számadó, S., & Takács, K. (2020). Scarce and directly beneficial reputations support cooperation. *Scientific reports*, 10(1), 1-12.
- Santos, F.P., Santos, F.C., & Pacheco, J.M. (2018). Social norm complexity and past reputations in the evolution of cooperation. *Nature*, 555(7695), 242-245.
- Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, 208, 104544.
- Sasson, N. J., Morrison, K. E., Pinkham, A. E., Faso, D. J., & Chmielewski, M. (2018). Brief report: Adults with autism are less accurate at predicting how their personality traits are evaluated by unfamiliar observers. *Journal of autism and developmental disorders*, 48, 2243-2248.
- Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management journal*, 43(6), 1248-1264.
- Schlenker, B. R., Chambers, J. R., & Le, B. M. (2012). Conservatives are happier than liberals, but why? Political ideology, personality, and life satisfaction. *Journal of Research in Personality*, 46(2), 127-146.

- Schlenker, B. R., & Leary, M. R. (1982). Social anxiety and self-presentation: A conceptualization model. *Psychological Bulletin*, 92, 641–669.
- Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology*, 43, 133-168.
- Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, 124(3), 325-333.
- Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: exploring the universal and culture-specific features of global self-esteem. *Journal of personality and social psychology*, 89(4), 623.
- Schoenmakers, S., Hilbe, C., Blasius, B., & Traulsen, A. (2014). Sanctions as honest signals—the evolution of pool punishment by public sanctioning institutions. *Journal of Theoretical Biology*, 356, 36-46.
- Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, 366(6466), eaau5141.
- Schwarz, N. (2012). Why researchers should think “real-time”: A cognitive rationale. *Handbook of research methods for studying daily life*, 22-42.
- Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness studies*, 4(1), 5-34.
- Scollon, C. N., Prieto, C. K., & Diener, E. (2009). Experience sampling: promises and pitfalls, strength and weaknesses. *Assessing well-being*, 157-180.
- Sebastián-Enesco, C., & Warneken, F. (2015). The shadow of the future: 5-year-olds, but not 3-year-olds, adjust their sharing in anticipation of reciprocation. *Journal of experimental child psychology*, 129, 40-54.
- Seinen, I., & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European economic review*, 50(3), 581-602.
- Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, 106(35), 15073-15078.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's  $\eta^2$ , a measure of local effect size, from PROC MIXED. *Frontiers in psychology*, 3, 111.

Semmann, D., Krambeck, H. J., & Milinski, M. (2004). Strategic investment in reputation. *Behavioral Ecology and Sociobiology*, 56(3), 248-252.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current directions in psychological science*, 24(2), 125-130.

Shaw, A., & Olson, K. (2015). Whose idea is it anyway? The importance of reputation in acknowledgement. *Developmental Science*, 18(3), 502–509.

Shepherd, D. A., & Zacharakis, A. (1999). Conjoint analysis: A new methodological approach for researching the decision policies of venture capitalists. *Venture Capital: An International Journal of Entrepreneurial Finance*, 1(3), 197-217.

Sherratt, T. N., & Roberts, G. (1998). The evolution of generosity and choosiness in cooperative exchanges. *Journal of theoretical biology*, 193(1), 167-177.

Shinada, M., & Yamagishi, T. (2007). Punishing free riders: Direct and indirect promotion of cooperation. *Evolution and human behavior*, 28(5), 330-339.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, 167, 201-211.

Silver, I. M., & Shaw, A. (2018). Pint-sized public relations: The development of reputation management. *Trends in Cognitive Sciences*, 22(4), 277–279.

Singh, M., & Boomsma, J. J. (2015). Policing and punishment across the domains of social evolution. *Oikos*, 124(8), 971-982.

[Singmann, H., Bolker, B., Westfall, J., & Aust, F. \(2016\). afex: Analysis of Factorial Experiments. R package version 0.16-1. https://CRAN.R-project.org/package=afex](https://CRAN.R-project.org/package=afex)

Singmann, H., Kellen, D., Cox, G. E., Chandramouli, S. H., Davis-Stober, C. P., Dunn, J. C., ... & Shiffrin, R. M. (2023). Statistics in the service of science: Don't let the tail wag the dog. *Computational Brain & Behavior*, 6(1), 64-83.

Sjåstad, H. (2019). Short-sighted greed? Focusing on the future promotes reputation-based generosity. *Judgment and Decision making*, 14(2), 199-213.

Smith, K. M., & Apicella, C. L. (2020). Partner choice in human evolution: The role of cooperation, foraging ability, and culture in Hadza campmate preferences. *Evolution and Human Behaviour*, 41(5), 354-366.

Smith, E. A., Bird, R. B., & Bird, D. W. (2003). The benefits of costly signaling: Meriam turtle hunters. *Behavioral Ecology*, 14(1), 116–126.

Smith, J. M., & Harper, D. (2003). *Animal signals*. Oxford University Press.

Smith, J. M., & Szathmary, E. (1997). *The major transitions in evolution*. OUP Oxford.

Snodgrass, S. E., Hecht, M. A., & Ploutz-Snyder, R. (1998). Interpersonal sensitivity: Expressivity or perceptivity?. *Journal of personality and social psychology*, 74(1), 238.

Spadaro, G., Tidli, I., Columbus, S., Jin, S., Ten Teije, A., CoDa Team, & Balliet, D. (2022). The Cooperation Databank: machine-readable science accelerates research synthesis. *Perspectives on Psychological Science*, 17(5), 1472-1489.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393.

Srivastava, S., & Beer, J. S. (2005). How self-evaluations relate to being liked by others: integrating sociometer and attachment perspectives. *Journal of personality and social psychology*, 89(6), 966.

[Srna, S., Barasch, A., & Small, D. \(2017\). When Modesty Pays. ACR North American Advances. <https://www.acrwebsite.org/volumes/1024799/volumes/v45/NA-45#>](https://www.acrwebsite.org/volumes/1024799/volumes/v45/NA-45#)

Steinmetz, J., Sezer, O., & Sedikides, C. (2017). Impression mismanagement: People as inept self-presenters. *Social and Personality Psychology Compass*, 11(6), e12321.

Stopfer, J. M., Egloff, B., Nestler, S., & Back, M. D. (2014). Personality expression and impression formation in online social networks: An integrative approach to understanding the processes of accuracy, impression management and meta-accuracy. *European Journal of Personality*, 28(1), 73-94.

Strimling, P., & Eriksson, K. (2014). Regulating the regulation: Norms about punishment. *Reward and punishment in social dilemmas*, 52-69.

Strohming, N., & Nichols, S. (2014). The essential moral self. *Cognition*, 131(1), 159-171.

- Stüber, R. (2020). The benefit of the doubt: willful ignorance and altruistic punishment. *Experimental Economics*, 23(3), 848-872.
- Sukhera, J., Miller, K., Scerbo, C., Milne, A., Lim, R., & Watling, C. (2020). Implicit stigma recognition and management for health professionals. *Academic Psychiatry*, 44, 59-63.
- Suknunan, S., & Bhana, A. (2022). Influence of employee-manager relationship on employee performance and productivity. *Problems and Perspectives in Management*, 20(3).
- Sun, B., Jin, L., Yue, G., & Ren, Z. (2023). Is a punisher always trustworthy? in-group punishment reduces trust. *Current Psychology*, 42(26), 22965-22975.
- Suzuki, S., & Akiyama, E. (2005). Reputation and the evolution of cooperation in sizable groups. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570), 1373–1377.
- Swickert, R., Hittner, J. B., Kitos, N., & Cox-Fuenzalida, L. E. (2004). Direct or indirect, that is the question: a re-evaluation of extraversion's influence on self-esteem. *Personality and Individual Differences*, 36(1), 207-217.
- Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology letters*, 6(5), 659-662.
- Sylwester, K., & Roberts, G. (2013). Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior*, 34(3), 201-206.
- Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, 6(3), 167.
- Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral ecology and sociobiology*, 65, 13-21.
- Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences*, 113(10), 2625-2630.
- Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8, 321–340.



Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annu. Rev. Psychol.*, 58, 345-372.

Tassone, A., & Foster, M. D. (2021). The Relationship between Dimensions of Collective Action, Introversion/Extroversion, and Collective Action Endorsement among Women. *Contention*, 9(1), 31-62.

Tennie, C., Frith, U., & Frith, C. D. (2010). Reputation management in the age of the world-wide web. *Trends in cognitive sciences*, 14(11), 482-488.

Teper, R., Zhong, C. B., & Inzlicht, M. (2015). How emotions shape moral behavior: Some answers (and questions) for the field of moral psychology. *Social and Personality Psychology Compass*, 9(1), 1-14.

Tetlock, P. E., Visser, P. S., Singh, R., Polifroni, M., Scott, A., Elson, S. B., ... & Rescober, P. (2007). People as intuitive prosecutors: The impact of social-control goals on attributions of responsibility. *Journal of experimental social psychology*, 43(2), 195-209.

Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021). Economic games: An introduction and guide for research. *Collabra: Psychology*, 7(1), 19004.

Tissera, H., Gazzard Kerr, L., Carlson, E. N., & Human, L. J. (2021). Social anxiety and liking: Towards understanding the role of metaperceptions in first impressions. *Journal of Personality and Social Psychology*, 121(4), 948.

Tjosvold, D. (1982). Effects of cooperative and competitive interdependence and task complexity on subordinates' productivity, perception of leader, and group development. *Canadian Journal of Behavioural Science*, 14(1), 24-34.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in cognitive sciences*, 12(12), 455-460.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5), 675-691.

Traulsen, A., Röhl, T., & Milinski, M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society B: Biological Sciences*, 279(1743), 3716-3721.

- Tsai, L. L., Trinh, M., & Liu, S. (2022). What makes anticorruption punishment popular? Individual-level evidence from China. *The Journal of Politics*, 84(1), 602-606.
- Tsankova, E., & Tair, E. (2021). Meta-accuracy of very first impressions: a mini review. *Frontiers in Psychology*, 12, 736534.
- Tsankova, E., & Tair, E. (2022). Meta-Accuracy on the Internet: Initial Tests of Underlying Dimensions, Contributing Factors, and Biases. *Frontiers in Psychology*, 13, 837931.
- Turner, R. E., Edgley, C., & Olmstead, G. (1975). Information control in conversations: Honesty is not always the best policy. *Kansas Journal of Sociology*, 69-89.
- Twardawski, M., Gollwitzer, M., Pohl, S., & Bošnjak, M. (2022). What drives second-and third-party punishment? Conceptual replications of the “intuitive retributivism” hypothesis. *Zeitschrift für Psychologie*, 230(2), 77.
- Tyler, T. R., & Boeckmann, R. J. (1997). Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. *Law and Society Review*, 237-265.
- Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, 10(1), 72-81.
- Ule, A., Schram, A., Riedl, A., & Cason, T. N. (2009). Indirect punishment and generosity toward strangers. *Science*, 326(5960), 1701-1704.
- Uskul, A. K., & Over, H. (2017). Culture, social interdependence, and ostracism. *Current Directions in Psychological Science*, 26(4), 371-376.
- Uskul, A. K., Kirchner-Häusler, A., Vignoles, V. L., Rodriguez-Bailon, R., Castillo, V. A., Cross, S. E., ... & Uchida, Y. (2023). Neither Eastern nor Western: Patterns of independence and interdependence in Mediterranean societies. *Journal of Personality and Social Psychology*, 125(3), 471-495.
- Uy, M. A., Foo, M. D., & Aguinis, H. (2010). Using experience sampling methodology to advance entrepreneurship theory and research. *Organizational Research Methods*, 13(1), 31-54.
- Vaish, A., Carpenter, M., & Tomasello, M. (2009). Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers. *Developmental psychology*, 45(2), 534.
- Vaish, A., Herrmann, E., Markmann, C., & Tomasello, M. (2016). Preschoolers value those who sanction non-cooperators. *Cognition*, 153, 43–51.

- Van de Calseide, P. P., Keren, G., & Zeelenberg, M. (2014). Decision time as information in judgment and choice. *Organizational Behavior and Human Decision Processes*, 125(2), 113-122.
- Van de Vliert, E., & Van Lange, P. A. (2019). Latitudinal psychology: an ecological perspective on creativity, aggression, happiness, and beyond. *Perspectives on Psychological Science*, 14(5), 860-884.
- van den Berg, P., Dewitte, S., & Wenseleers, T. (2021). Uncertainty causes humans to use social heuristics and to cooperate more: An experiment among Belgian university students. *Evolution and Human Behavior*, 42(3), 223-229.
- van Dijk, E., & De Dreu, C. K. (2021). Experimental games and social decision making. *Annual Review of Psychology*, 72, 415-438.
- Van Vugt, M., Roberts, G., & Hardy, C. (2007). Competitive altruism: Development of reputation-based cooperation in groups. *Handbook of evolutionary psychology*, 531-540.
- Vaughan-Johnston, T. I., MacGregor, K. E., Fabrigar, L. R., Evraire, L. E., & Wasylikiw, L. (2021). Extraversion as a moderator of the efficacy of self-esteem maintenance strategies. *Personality and Social Psychology Bulletin*, 47(1), 131-145.
- Villatoro, D., Andrighetto, G., Brandts, J., Nardin, L. G., Sabater-Mir, J., & Conte, R. (2014). The norm-signaling effects of group punishment: combining agent-based simulation and laboratory experiments. *Social science computer review*, 32(3), 334-353.
- von Rueden, C. R., Gurven, M., & Guala, F. (2012). When the strong punish: Why net costs of punishment are often negligible. *Behavioral and Brain Sciences*, 35(1), 43.
- von Rueden, C. R., Redhead, D., O'Gorman, R., Kaplan, H., & Gurven, M. (2019). The dynamics of men's cooperation and social status in a small-scale society. *Proceedings of the Royal Society B*, 286(1908), 20191367.
- Vonasch, A. J., Reynolds, T., Winegard, B. M., & Baumeister, R. F. (2018). Death before dishonor: Incurring costs to protect moral reputation. *Social Psychological and Personality Science*, 9(5), 604-613.
- Wainwright, P. E., Leatherdale, S. T., & Dubin, J. A. (2007). Advantages of mixed effects models over traditional ANOVA models in developmental studies: a worked example in a

mouse model of fetal alcohol syndrome. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 49(7), 664-674.

Walker, L. J., & Hennig, K. H. (2004). Differing conceptions of moral exemplarity: just, brave, and caring. *Journal of personality and social psychology*, 86(4), 629.

Wang, Y., Zheng, H., Tang, W., Liu, B., & Zhang, Z. (2024). Doing good is not equal to good people: The wealth-based gap in prosocial motive attributions. *Asian Journal of Social Psychology*.

Warneken, F., & Tomasello, M. (2007). Helping and cooperation at 14 months of age. *Infancy*, 11(3), 271-294.

Warneken, F., Sebastián-Enesco, C., Benjamin, N. E., & Pieloch, K. A. (2019). Pay to play: Children's emerging ability to use acts of generosity for selfish ends. *Journal of Experimental Child Psychology*, 188, 104675.

Wason, K. D., Polonsky, M. J., & Hyman, M. R. (2002). Designing vignette studies in marketing. *Australasian Marketing Journal*, 10(3), 41-58.

Watling, D. (2019). *Reputation management*. The Encyclopedia of Child and Adolescent Development, 1-10.

Watling, D., & Banerjee, R. (2012). Children's understanding of disclaimers. *Social Cognition*, 30(1), 18-36.

Watson, D., Hubbard, B., & Wiese, D. (2000). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of personality and social psychology*, 78(3), 546.

Waytz, Adam, Ravi Iyer, Liane Young, Jonathan Haidt, and Jesse Graham. "Ideological differences in the expanse of the moral circle." *Nature communications* 10, no. 1 (2019): 1-12.

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current biology*, 12(12), 1012-1015.

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, 288(5467), 850-852.

- Weiner, B., Osborne, D., & Rudolph, U. (2011). An attributional analysis of reactions to poverty: The political ideology of the giver and the perceived morality of the receiver. *Personality and Social Psychology Review*, 15(2), 199-213.
- Weiss, A., Michels, C., Burgmer, P., Mussweiler, T., Ockenfels, A., & Hofmann, W. (2021). Trust in everyday life. *Journal of Personality and Social Psychology*, 121(1), 95.
- Weiss, B., & Feldman, R. S. (2006). Looking good and lying to do it: Deception as an impression management strategy in job interviews. *Journal of Applied Social Psychology*, 36(4), 1070-1086.
- Welker, C., Walker, J., Boothby, E., & Gilovich, T. (2023). Pessimistic assessments of ability in informal conversation. *Journal of Applied Social Psychology*, 53(7), 555-569.
- Wellman, H. M. (1992). *The child's theory of mind*. The MIT Press series in learning, development, and conceptual change.
- West, S. A., Griffin, A. S., & Gardner, A. (2007). Evolutionary explanations for cooperation. *Current biology*, 17(16), R661-R672.
- West, S. J., Parton, D. M., & Chester, D. (2022). Harming in order to help: an empirical demonstration of prosocial aggression. *PsyArXiv*.
- West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological review*, 118(2), 357.
- West, B. T., Welch, K. B., & Galecki, A. T. (2022). *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC.
- West-Eberhard, M. J. (1983). Sexual selection, social competition, and speciation. *The quarterly review of biology*, 58(2), 155-183.
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour?. *Journal of animal ecology*, 75(5), 1182-1189.
- Wickham, H., & Girlich, M. (2022). tidyr: Tidy Messy Data. R package version 1.2.0. <https://CRAN.R-project.org/package=tidyr>.
- Wiessner, P. (2005). Norm enforcement among the Ju/'hoansi Bushmen: A case of strong reciprocity?. *Human Nature*, 16, 115-145.

- Wiessner, P. (2009). Experimental games and games of life among the Ju/'hoan Bushmen. *Current Anthropology*, 50(1), 133-138.
- Wiessner, P. (2020). The role of third parties in norm enforcement in customary courts among the Enga of Papua New Guinea. *Proceedings of the National Academy of Sciences*, 117(51), 32320-32328.
- Wilks, T. (2004). The use of vignettes in qualitative research into social work values. *Qualitative social work*, 3(1), 78-87.
- Williams, K. D. (2007). Ostracism. *Annual Review of Psychology*, 58, 425–452.
- Williams, K. D., Cheung, C. K., & Choi, W. (2000). Cyberostracism: effects of being ignored over the Internet. *Journal of personality and social psychology*, 79(5), 748.
- Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European review of social psychology*, 16(1), 155-188.
- Wojciszke, B., Bazinska, R., & Jaworski, M. (1998a). On the dominance of moral categories in impression formation. *Personality and social psychology bulletin*, 24(12), 1251-1263.
- Wojciszke, B., Dowhyluk, M., & Jaworski, M. (1998b). Moral competence-related traits: how do they differ?. *Polish Psychological Bulletin*.
- Wolf, W., Nafe, A., & Tomasello, M. (2021). The development of the liking gap: Children older than 5 years think that partners evaluate them less positively than they evaluate their partners. *Psychological Science*, 32(5), 789–798.
- Wood, D., & Furr, R. M. (2016). The correlates of similarity estimates are often misleadingly positive: The nature and scope of the problem, and some solutions. *Personality and Social Psychology Review*, 20(2), 79-99.
- Wu, J., Balliet, D., & Van Lange, P. A. (2015). Reputation management: Why and how gossip enhances generosity. *Evolution and Human Behavior*, 37(3), 193-201.
- Wu, J., Balliet, D., & Van Lange, P.A. (2016a). Gossip versus punishment: the efficiency of reputation to promote and maintain cooperation. *Scientific Reports*, 6(1), 23919.
- Wu, J., Balliet, D., & Van Lange, P.A. (2016b). Reputation, gossip, and human cooperation. *Social and Personality Psychology Compass*, 10(6), 350–364.

Wu, J., Luan, S., & Raihani, N. (2022). Reward, punishment, and prosocial behavior: Recent developments and implications. *Current Opinion in Psychology*, 44, 117-123.

Wu, J. J., Zhang, B. Y., Zhou, Z. X., He, Q. Q., Zheng, X. D., Cressman, R., & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences*, 106(41), 17448-17451.

Wu, T., & Zheng, Y. (2019). Is impression management through status updates successful? Meta-accuracy and judgment accuracy of big five personality traits based on status updates from social network sites in China. *Frontiers in psychology*, 10, 443501.

Xiao, E. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, 77(1), 321-344.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology*, 51(1), 110.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., ... & Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*, 109(50), 20364-20368.

Yamagishi, T., Terai, S., Kiyonari, T., Mifune, N., & Kanazawa, S. (2007). The social exchange heuristic: Managing errors in social exchange. *Rationality and Society*, 19, 259–291.

Yang, S., & Chen, S. (2022). Normative misperception in third-party punishment: An explanation from the perspective of belief in a just world. *Acta Psychologica Sinica*, 54(3), 281.

Yoeli, E., Hoffman, M., Rand, D. G., & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences*, 110(2), 10424-10429.

Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of cognitive neuroscience*, 21(7), 1396-1405.

Yuan, Z., & Yang, Y. (2005). Combining linear regression models: When and how?. *Journal of the American Statistical Association*, 100(472), 1202-1214.

Yudkin, D. A., Van Bavel, J. J., & Rhodes, M. (2020). Young children police group members at personal cost. *Journal of Experimental Psychology: General*, 149(1), 18.

- Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, 2(3), 7–10. Retrieved from <https://CRAN.R-project.org/doc/Rnews/>
- Zheng, X., & Nie, P. (2013). Effective punishment needs legitimacy. *Economic Record*, 89(287), 522-544.
- Zhu, Q., Lo, L. Y. H., Xia, M., Chen, Z., & Ma, X. (2022). Bias-aware design for informed decisions: Raising awareness of self-selection bias in user ratings and reviews. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1-31.
- Zinko, R., Ferris, G. R., Blass, F. R., & Laird, M. D. (2007). Toward a theory of reputation in organizations. *Research in Personnel and Human Resources Management*, 26, 163–204.
- Zinko, R., Furner, Z. Z., Hunt, J., & Dalton, A. (2017). Establishing a reputation. *Journal of Employment Counseling*, 54(2), 87-96.
- Zuckerman, M., & Kuhlman, D. M. (2000). Personality and risk-taking: common bisocial factors. *Journal of personality*, 68(6), 999-1029.
- Zuur, A. F., & Ieno, E. N. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, 7(6), 636-645.
- Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, 1(1), 3-14.