

**Mapping the Moral Mind:
Computational & Formal Approaches to
Understanding Prosocial Behaviour**

Maximilian Maier

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Experimental Psychology
University College London

July 31, 2025

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

Abstract

Many challenges in the 21st century (e.g., addressing climate change) are related to (failures of) moral and prosocial decision-making. In this thesis, I investigate how humans make these decisions, drawing on data from over 25 million individuals (5,934 from experimental studies and the rest from meta-analyses) and tackle two key methodological bottlenecks: publication bias and (lack of) computational modelling.

In Part I (Chapters 1–3), I use novel statistical techniques, which I developed to adjust for publication bias in meta-analysis, to critically evaluate the evidence for Construal Level Theory, the Identifiable Victim Effect, and nudging. I document strong publication bias in all three of these areas, with meta-analytic mean effects not distinguishable from zero after adjusting for this bias.

Part II of the thesis offers a more constructive counterpoint by developing novel paradigms and computational models. Chapter 4 examines scope insensitivity, evaluates the Unit Asking Method as an intervention to address it, and proposes an improved Sequential Unit Asking technique. Chapter 5 investigates decisions under extinction risk (e.g., death or human extinction) in individual and collective decisions using a novel experimental paradigm. I derive optimal strategies for these types of decisions and use (dependent) mixture modelling to describe individual-level behaviour. I find that, while participants are relatively good in terms of the qualitative strategies employed, their decisions are nevertheless affected by behavioural biases. Chapter 6 studies how people learn to make moral decisions using a novel reinforcement learning paradigm. Drawing on research on strategy selection learning and comparing different reinforcement learning models, I show that

learning about strategies (moral rules or cost-benefit reasoning) rather than specific behaviours predicts generalisation to incentive-compatible donation decisions.

Overall, this thesis demonstrates limitations in existing approaches while highlighting how formal and computational modelling, and publication-bias-adjusted meta-analysis can advance the understanding of moral and prosocial decision-making.

Impact Statement

This thesis investigates how humans make moral and prosocial decisions, drawing on data from over 25 million individuals. It contributes new statistical methods for bias-corrected meta-analysis and novel experimental paradigms and computational models for understanding how people respond to societal challenges. These contributions have the potential to benefit both academic research and decision making beyond academia.

With regards to academic research, the bias-adjustment techniques I developed offer a generalisable and scalable approach to re-evaluating evidence across disciplines that rely on meta-analysis, such as psychology, economics, and medicine. By applying these tools to influential domains such as Construal Level Theory, nudging, and the Identifiable Victim Effect, I show that some widely accepted findings are contaminated by publication bias, which is important knowledge for researchers working on these topics and underscores the need to adopt research practices that mitigate bias.

In addition, the computational models and novel experimental paradigms introduced in Part II can serve as reusable tools for other researchers studying these topics (to facilitate this, I make all data, experimental, and analysis code publicly available). I hope that this will bring more research attention to important domains that had previously been neglected. The academic impact of this thesis is also underscored by the fact that many of the chapters have been published or are in press in top disciplinary or multidisciplinary journals, such as *PNAS*, *Nature Human Behaviour*, and *Cognitive Psychology*.

The findings in this thesis are not only of interest to other academics but also to policymakers and other stakeholders. First, the strong publication bias in research on nudging and the identifiable victim effect shows that many behavioural interventions that are used by governments and NGOs are less effective than previously thought. To address this, Chapter 3 highlights possible improvements in research practices by Nudge Units (government agencies or independent companies that de-

velop and test behavioural interventions) to address these challenges, and Chapter 4 develops an effective intervention (Sequential Unit Asking) that can be used by charities and NGOs for fundraising.

Further, my research on decisions under extinction risk offers a framework for understanding how individuals and groups perceive and act on long-term threats. I also use this paradigm to test several interventions that can reduce risk-taking under extinction risk. These insights can inform public communication strategies and global risk governance.

Finally, the reinforcement learning paradigm developed in Chapter 6 shows that when people receive feedback about the consequences of their decisions, they adopt decision strategies that benefit the greater good. These insights can be used to design behavioural interventions and educational programs aimed at fostering more impartial prosocial decision making, by prompting people to reflect on the direct and indirect consequences of their moral decisions.

Articles & Preprints

The chapters of this thesis are based on the following manuscripts. Detailed author contributions are provided at the end of each chapter.

Chapter 2: Identifiable Victim Effect

Maier, M.*, Wong, Y.*, & Feldman, G.* (2023). Revisiting and rethinking the identifiable victim effect: Replication and extension of Small, Loewenstein, and Slovic (2007). *Collabra: Psychology*, 9(1). <https://doi.org/10.1525/collabra.90203>.

Chapter 3: Nudging

Maier, M.*, Bartoš, F.*, Stanley, T.D., Shanks, D.R., Harris, A.J.L., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, 119(31), e2200300119. <https://doi.org/10.1073/pnas.2200300119>

Maier, M.*, Bartoš, F.*, Raihani, N., Shanks, D.R., Stanley, T.D., Wagenmakers, E.-J., & Harris, A.J.L. (2024). Exploring open science practices in behavioural public policy research. *Royal Society Open Science*, 11. <http://doi.org/10.1098/rsos.231486>.

Chapter 4: Scope Insensitivity & (Sequential) Unit Asking

Maier, M., Caviola, L., Schubert, S., & Harris, A.J.L. (2023). Investigating (Sequential) Unit Asking – An Unsuccessful Quest for Scope Sensitivity in WTD Judgements. *Journal of Behavioural Decision Making*, 36(4), e2335. <https://doi.org/10.1002/bdm.2335>.

Chapter 5: Decisions Under Extinction Risk

Maier, M., Harris, A. J. L., Kellen, D., & Singmann, H. (in press). Decision Making Under Extinction Risk. *Cognitive Psychology*. <https://doi.org/10.31234/osf.io/qjd37>

Maier, M.*, Pilgrim, C.*, Mann R.P., & Singmann, H. (under review). Collective Decision Making Under Extinction Risk [Stage 1 registered report].

Chapter 6: Moral Learning

Maier, M.*, Cheung, V.*, & Lieder, F. (in press). Learning from Outcomes Shapes Reliance on Moral Rules versus Cost-Benefit Reasoning. *Nature Human Behaviour*. <https://doi.org/10.31234/osf.io/gjf3h>

Tahmasebi, Z.*, **Maier, M.***, Cheung, V.*, Cushman, F., & Lieder, F. (in press). Disentangling Model-Based and Model-Free Moral Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*. https://doi.org/10.31234/osf.io/azhnq_v1

Other relevant work

In addition to these papers, I conducted further research during my PhD, which is not included in this thesis due to space and topical constraints. I include some of the most relevant work that underlies the thinking and methodology in this thesis below. For a full list of my publications, see my Google Scholar profile <https://scholar.google.com/citations?user=JInh9X8AAAAJ&hl=en>.

Cheung, V.*, **Maier, M.***, & Lieder, F. (in press). Large language models amplify human biases in moral decision-making. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.31234/osf.io/aj46b>

Bartoš, F.*, **Maier, M.***, Stanley, T. D., & Wagenmakers, E.-J. (in press). Robust Bayesian meta-regression – Model-averaged moderation analysis in the presence of publication bias. *Psychological Methods*. <https://doi.org/10.31234/osf.io/98xb5>.

Bartoš, F.*, **Maier, M.***, Wagenmakers, E.-J., Nippold, F., Doucouliagos, H., Ioannidis, J.P.A., Otte, W.M., Sladekova, M., Deressa, T.K., Bruns, S.B, Fanelli, D.,

& Stanley, T.D. (2024) Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics. *Research Synthesis Methods*, 5(3), 500-511. <https://doi.org/10.1002/jrsm.1703>.

Maier, M. van Dongen, N., & Borsboom, D. (2024). Comparing theories with the Ising model of explanatory coherence. *Psychological Methods*. 29(3), 519-536. <https://doi.org/10.1037/met0000543>.

Maier, M.*, Bartoš, F.*, & Wagenmakers, E.-J. (2023). Robust Bayesian Meta-Analysis: Addressing publication bias with model-averaging. *Psychological Methods*, 28(1), 107-122. <https://doi.org/10.1037/met0000405>.

Bartoš, F., **Maier, M.**, & Wagenmakers, E.-J.(2023). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, 14, 99-116. <https://doi-org.libproxy.ucl.ac.uk/10.1002/jrsm.1594>.

Bartoš, F.*, **Maier, M.***, Shanks, D.R., Stanley, T.D., Sladekova, M., & Wagenmakers, E.-J. (2022). Meta-analyses in psychology often overestimate evidence for and size of effects. *Royal Society Open Science*, 10 (230224). <https://doi.org/10.1098/rsos.230224>.

* indicates that authors contributed equally.

Acknowledgements

I have just one wish for you — the good luck to be somewhere where you are free to maintain the kind of integrity I have described, and where you do not feel forced by a need to maintain your position in the organization, or financial support, or so on, to lose your integrity. May you have that freedom.

— Feynman, 1985 (p.234)

I feel incredibly fortunate that during my research career so far, I was always working in environments that allowed me to focus on the unbiased pursuit of truth, being free to work on whichever topics I consider most interesting and important (and even getting paid for doing so). This would not have been possible without the support of many people who have helped and inspired me.

I am most grateful to my supervisors, Adam Harris and Henrik Singmann, for encouraging me to pursue my own research interests while providing consistent support and thoughtful feedback throughout the PhD. Both of them were incredibly generous with their time and ideas, and supported not only the work in this thesis but my development as a researcher more generally. I learned much more from them than I can mention in this acknowledgement section, but I want to particularly thank Adam for teaching me how to best go about phrasing critical pieces, such as reanalyses of others' research, and Henrik for teaching me how to write and evaluate custom models to model behavioural data in Stan, both skills that I am confident will benefit me throughout my academic career.

Next to my supervisors, I was fortunate to meet many exceptional mentors and collaborators, without whom the work in this thesis would not have been possible;

they have made my academic journey thus far productive and fun. Although I have done my best to thank everyone who has helped me, memory is fallible; if your name is missing, please accept my apologies and know that your support was very much appreciated.

First, I want to thank Daniël Lakens and members of his lab (Peder Isager, Anne Scheel, Leo Tiokhin), who hosted me for a research visit back during my undergraduate, even though I had no research experience and was still doing my statistics using Microsoft Excel. Daniël and his lab were exceptionally generous with their time (and money buying me drinks), while allowing me to work on whatever I was most interested in and found most educational. My time in Eindhoven led me to decide to pursue research as a career, and it is possible that I would be working in a completely different profession without this experience.

Second, I thank everyone who supported me during my time at the University of Amsterdam and the Research Master's Psychology more generally. This master's programme taught me the technical skills and research philosophy that guide much of my work to this day. I thank Denny Borsboom and Noah van Dongen, who taught me the value of theory development and the importance of implementing theories in the form of computational models and more generally, how much fun theoretical work can be (our 90 minute weekly meetings discussing philosophy of science were always the highlight of my week during the Covid lockdown).

I would like to thank Eric-Jan Wagenmakers for supporting my research ideas and plans when I was a research master's student with little experience, and teaching me the value of Bayesian statistics and Open Science, as well as polishing up my writing and teaching me much about how to write a compelling methodological paper. I thank František Bartoš for all his contributions to our joint projects and his hard work on the RoBMA R-package, without which many of the reanalyses included here would not have been possible. For contribution to my meta-analysis research, I also thank Tom Stanley, who was an invaluable collaborator throughout and invited me to my first in-person conference, as well as Chris Doucouliagos, John Ioannidis, and Maya Mathur.

After my master's, I took some time to reflect on the most important research directions in psychology, which motivated me to pivot to experimental research and to move to UCL and pursue a PhD on the psychology of existential risks and moral and prosocial decision making more broadly. I want to thank everyone whom I consulted during this time, including Lucius Caviola, 80,000 Hours (in particular Michelle Hutchinson), and of course my friends and family, who all provided valuable advice. Most importantly, I would like to thank Longview Philanthropy for funding my PhD and making my shift in research focus possible, which allowed me to pursue the projects I believed to be most impactful.

At UCL, next to my supervisors, I want to especially thank David Shanks, who encouraged me to continue some of my investigations into publication bias in various subfields of psychology, provided mentorship and advice, and supported me with grant and fellowship applications. I further want to thank David Lagnado, Nichola Raihani, Stephen Dewitt, and Maarten Speekenbrink, who supported me in various ways, including collaboration on projects, and feedback on lab presentations and job talks, and my research assistants Leah Wu, Megan Oh, and Talia Ozluk, who provided valuable support for some of the work in this thesis and other research. I also want to thank Charlie Pilgrim for his collaboration on the collective extinction gambling tasks, and all the other PhD students and postdocs who provided much helpful feedback and collaboration and made the PhD experience fun: Harry Coulson, Helen Qiao, Calvin Deans-Browne, Greta Sanna, Tianshu Chen, Victor Btesh, Tris Papakonstantinou, Kevin O'Neill, Caro Echterbeck, and Ram Mungur.

Outside UCL, many other collaborators supported me during my PhD. In particular, I would like to thank Falk Lieder for inspiring my thinking on how to integrate my interests in computational modelling with my interests in moral and prosocial decision-making, as well as for his mentorship and advice more generally. Next to his time, I also want to thank Falk for his generous financial support, paying for all the studies in Chapter 6 of this thesis at a combined cost of well-above £15 000. In addition to Falk, I also thank other members of the Rational Altruism lab, in

particular Glen Spiteri and Zahra Thamasebi, who helped with some of the projects included here. I also want to thank Gilad Feldman for his collaboration, support, and encouragement for my work on open science and meta-science, and more generally, his hard work replicating many of the key findings in JDM on which this thesis builds. Further, I want to thank Tom Griffiths for hosting me for a research visit during my PhD and members of Tom's lab, including Bonan Zhao, Jianqiao Zhu, and Logan Nelson, for making me feel welcome during my stay.

Lastly, I would like to thank my family and friends. In particular, I want to thank Erdem, Patrick, René, Jens, and Jonas for advising on the various decisions that I needed to make during the PhD and for listening to any problems encountered. I thank Jonas Wolfram in particular for often providing valuable feedback on pilot studies from the perspective of someone who is not trained in psychological research. He prevented more than one incomprehensible study from going out to participants. I also want to thank Vanessa for all the support and help during the PhD and for always being there for me. Thank you for providing feedback on draft versions of almost all the papers included here and for our collaborations, which were the projects of my PhD that I enjoyed the most. More importantly, thank you for all the time we spent together and the happiness you brought into my life. Meeting you was certainly the best part of my PhD journey, and I am excited for everything that's still ahead. Finally, I want to thank my sister and brother-in-law, as well as my parents, who for sure made the biggest difference to my life and trajectory, and supported my interests in research, science, and psychology from a young age. Thank you for always believing in me, for your reliable support, and for your genuine interest in and enthusiasm for what I do.

Contents

0	Introduction	25
0.1	Two Key Issues in Research on Moral and Prosocial Decision Making: Publication Bias and Lack of Formal Modelling.	27
0.2	Overview of the Thesis	29
I	Revisiting Classic Theories and Interventions with Novel Publication Bias Adjustment Methods	32
1	Construal Level Theory	35
1.1	Motivation for the Current Research	36
1.2	Reanalysing a Comprehensive Meta-Analysis on Construal Level Theory	38
1.3	Sequential Robust Bayesian Meta-Analysis of Recent Studies	45
1.4	Quantitative Assessment of Publication in a Broader Set of Studies using Z-Curve	47
1.5	Conclusion	49
2	Identifiable Victim Effect	53
2.1	Reanalysis of a Meta-Analysis on the Identifiable Victim Effect Suggests the Need to Revisit the Phenomenon	54
2.2	Replication of Small et al. (2007)	57
2.3	Discussion	65

3	Nudging	71
3.1	No Evidence for Nudging After Adjusting for Publication Bias	71
3.2	Exploring Open Science Practices in Behavioural Public Policy Research	75
3.2.1	Exploring Potential Reporting Biases in Nudge Unit Trials Using Bias Correction Techniques	77
3.2.2	Pre-analysis Plans Leave Scope for Selective Reporting . . .	81
3.2.3	Evidence Based Public Policy Needs to Increase Transparency	84
 II Experimental & Computational Modelling Work to Address Neglected Issues in Moral and Prosocial Decision Making		 88
4	Scope Insensitivity & (Sequential) Unit Asking	90
4.1	Unit Asking	91
4.2	Extending Unit Asking to <i>Sequential</i> Unit Asking	91
4.3	Does (Sequential) Unit Asking Make People Scope Sensitive? . . .	92
4.4	Study 1a	94
4.5	Study 1b	102
4.6	Study 2	105
4.7	Study 3	112
4.8	Study 4	119
4.9	General Discussion	125
5	Decisions Under Extinction Risk	130
5.1	Individual Decision Making Under Extinction Risk	131
5.1.1	Towards an Extinction Gambling Task (EGT)	134
5.1.2	Experiment 1: Empirical Choice Patterns for the Lose and Keep Conditions With 100 Trials	141
5.1.3	Experiment 2: Introducing a Reset Condition	152

5.1.4	Experiment 3: What Psychological Factors Shape Decisions Under Extinction Risk?	159
5.1.5	General Discussion	170
5.2	Collective Decision Making Under Extinction Risk (Stage 1: Registered Report with Pilot Data)	177
5.2.1	Method	184
5.2.2	Pilot Results	190
5.2.3	Conclusion	192
6	Moral Learning	194
6.1	Metacognitive Moral Learning in Realistic Moral Dilemmas	194
6.1.1	A Theory of Metacognitive Moral Learning	198
6.1.2	A New Experimental Paradigm Using Realistic Trolley-Type Dilemmas with Outcomes	200
6.1.3	Computational Models of Moral Learning from Consequences	203
6.1.4	Experiment 1	209
6.1.5	Experiment 2	214
6.1.6	Discussion	221
6.2	Disentangling Model-Based and Model-Free Moral Learning	229
6.2.1	The Moral Dilemma Two-Step Task	230
6.2.2	Experiment	232
6.2.3	Computational Modeling	238
6.2.4	Discussion	241
7	General Discussion	245
7.1	Domain Insights About Moral and Prosocial Decision Making	246
7.1.1	Insights from Publication Bias Adjusted Meta-Analysis	246
7.1.2	Insights from Experimental Work	247
7.2	Implications for Methodology and Research Practices	250
7.2.1	Mitigating Publication Bias	250
7.2.2	Benefits of Computational Modelling	251

7.3	Where Next? Bridging Publication Bias Adjusted Meta-Analysis and Computational Modelling	252
7.4	Conclusion	253
Appendices		296
A Construal Level Theory		296
A.1	Additional RoBMA Results	296
A.2	Coding Procedure RoBMA	296
A.3	Coding Procedure Z-curve	298
B Sequential Unit Asking		300
B.1	Pilot 2 - Budget Constrains as Explanation for The Effectiveness of SUA over CUA	300
B.2	Pilot 3 - Priming Scope Sensitivity	303
B.3	Prior Predictives for BRMs model	305
B.4	Share of Participants Donating for DA, CUA, and SUA conditions in The Different Studies	305
B.5	Step-by-Step Guide to the Bayesian Analysis	305
C Decisions Under Extinction Risk		310
C.1	Existing Paradigms and Their Limitations	310
C.2	Expected Value vs. Expected Utility Optimal Strategies for the Three Conditions	312
C.3	A Priori (Non-Dynamic) Optimal Strategy for the Lose Condition .	312
C.4	Estimating False Positive Rate and Power	313
C.5	Full Model Specification and Implementation	315
C.6	Heatmaps of Optimal Choice as a Function of Endowment and Trial Number	317
C.7	Comparison of Posterior Predictions and Data	319
C.8	Additional Results	320
C.9	Optimal Strategies Collective Game	323

C.10 Details on Estimated Average Riskiness, False Positive Rate, and
Power for the Collective Task 328

List of Figures

1.1	Research on Construal Level Theory Is Increasing Rapidly	36
1.2	Funnel Plots for Soderberg et al. (2015)	39
1.3	Footprints of Publication Bias in the Soderberg et al. (2015)	43
1.4	Z-Curve Analysis of Test-Statistics from Construal Level Theory Studies	49
2.1	Footprint of Publication Bias in Lee and Feeley (2016)	56
2.2	Hypothetical Donations: Interaction of Identifiability and Explicit Learning	63
3.1	Correcting for Publication Bias Suggests No Evidence for the Mean Effect of Nudging	73
3.2	Distribution of All Effect Sizes and Visualization of the Meta- Analytic Models.	82
3.3	Aspects of Pre-Analysis Plans of Trials Run by Nudge Units.	83
4.1	Densities of WTD Judgements in Study 1a After Winsorizing.	98
4.2	Median WTDs for Each Step in Study 1a.	101
4.3	WTD Distribution for the Full Scope of 100 in Study 1a.	101
4.4	Median WTDs for Each Step in Study 1b.	104
4.5	WTD Distribution for the Full Scope in Study 1b.	104
4.6	Median WTDs for Each Step in Study 2.	109
4.7	WTD Distribution for the Full Scope in Study 2.	109
4.8	Median WTDs for Each Step in Study 3.	115
4.9	WTD Distribution for the Full Scope in Study 3.	116

4.10	Median Contingency Valuation Judgements for Each Step in Study 4.	122
4.11	Contingency Valuation Distribution for the Full Scope in Study 4.	122
5.1	The Experimental Setup as Shown to Participants on the Instruction Screen	143
5.2	Risky Choices and Survival Times in Experiment 1	145
5.3	Six Different Strategies Identified by the (Dependent) Mixture Model	149
5.4	Proportion of Participants Allocated to Each of the Six Strategies in Experiment 1	151
5.5	Parameter Estimates from the Computational Models in Experiment 1	151
5.6	Risky Choices and Survival Times in Experiment 2	156
5.7	Proportion of Participants Allocated to Each of the Six Strategies in Experiment 2	158
5.8	Participants' Switch Points and Optimal Switch Points for Experi- ment 2	159
5.9	Risky Choices and Survival Times for Experiment 3a	162
5.10	Risky Choices for Each Maximum Trial Number and Extinction Condition for Experiment 3b	167
5.11	Illustration of the Median Voting Condition as Shown to Participants	187
5.12	Annotated Task Setup as Shown to Participants in the Collective Conditions	188
5.13	Estimated Average Riskiness and Per Trial Proportion of Risky Choices	191
6.1	Meta-Control of Moral Decision-Making is Informed by Learning from Previous Decisions.	199
6.2	The Moral Learning Paradigm in Experiment 1.	201
6.3	Learning from Consequences Shapes Reliance on Moral Rules Ver- sus Cost-Benefit Reasoning.	213
6.4	Endorsement of Sacrificial Harm and Deontology Are Moderated By Evidence for Metacognitive Learning in Experiment 1.	214

6.5	Donations to CBR Charities Are Moderated By Evidence for Metacognitive Learning in Experiment 1.	215
6.6	Metacognitive Learning Transfers to Measures of Moral Convictions and Donation Decisions in Experiment 2.	220
6.7	The Moral Dilemma Two-Step Task	233
6.8	Participants' Choices Reflect Both Model-Based and Model-Free Learning	237
7.1	Where Next? Bridging Meta-Analytic Work and Computational Modelling	254
B.1	Prior Predictives for the Lognormal Model.	306
C.1	Optimal Strategies According to Expected Value vs. Expected Utility	312
C.2	Heatmaps Visualizing the Optimal Choice As a Function of Trial Number and Money for Experiment 1 - Lose Condition.	317
C.3	Heatmap Visualizing the Optimal Choice As a Function of Trial Number and Money for Experiment 2 - Reset Condition.	318
C.4	Model Predictions vs. Data in the Lose Condition (Exp. 1)	319
C.5	Model Predictions vs. Data in the Keep Condition (Exp. 1)	319
C.6	Number of Risky Choices, Proportion of Risky Choices, Trial Number of First Risky Choice, and Money Earned Between Conditions for Experiment 1	320
C.8	Introducing Endowment and Losses Does Not Affect the Strategies Participants Employ	321
C.7	Beta Coefficients of Those Participants Assigned to the Gradual Strategy.	321
C.9	Proportion of Strategies for the Two Different Maximum Scopes in the Keep condition	322
C.10	Proportion of Strategies for the Two Different Maximum Scopes in the Lose condition	322

C.11 Switch Points in Comparison to the Optimal Switch Point in the
Keep Condition 323

C.12 Model Predictions (Red) vs Data for the Pilot Study 331

List of Tables

1.1	Subgroup Analysis based on Significant Moderators in Soderberg et al. (2015)	52
2.1	Replication and Extension: Experimental Design	59
2.2	Hypothetical Donations: Statistical Tests for Identifiability and Explicit Learning	64
3.1	Comparison of Unadjusted and Adjusted Effect Size Estimates for All Studies and for Subsets of Studies Based on Different Categories or Domains.	74
4.1	Share of Participants Showing a Strong Monotonic Increase in Scope 10,000 SUAI (Increase per Step Constant) Condition and Scope 100 SUA Condition.	111
4.2	Summary of Results Across All Four Studies	126
5.1	Design Table	183
6.1	Cognitive Modeling Results Showing the Proportions of Participants that Relied on Each Type of Learning in Experiments 1-2.	218
6.2	Cognitive Modeling Results Showing the Proportions of Participants that Relied on Each Learning Mechanism in Experiments 1-2.	219
A.1	Summary of RoBMA Estimates for Soderberg et al. (2015).	296
A.2	Summary of RoBMA Estimates for Soderberg et al. (2015) with $ \delta < 1.5$	297

A.3	Summary of RoBMA Estimates for Soderberg et al. (2015) When Using Only Selection Models.	297
A.4	Summary of RoBMA Estimates for Soderberg et al. (2015) When Using Only Selection Models with $ \delta < 1.5$	298
B.1	Median WTD judgements For The Seven Conditions of Study 2 . . .	305
B.2	Median WTD judgements For The Seven Conditions of Study 3 . . .	307
C.1	Power Analysis for Mixed Model and T-Testing Proportions	314
C.2	Possible Combinations of Survival Outcomes and Their Associated Probabilities for Five Players, which Play four Risk Choices	324
C.3	Comparison Between Different Optimal Strategies in Terms of Pay- offs and Proportion Risky	326
C.4	Individually Optimal Strategy in the Collective Game	327
C.5	Influence of Selection Effects on False Positive Rate (FPR)	330

Chapter 0

Introduction

The PlayPump — a merry-go-round that pumps water while children play on it — offers a vivid illustration of the limitations in human moral and prosocial decision making. Inspired by the potential to pump fresh water without the heavy work of manually operating the pump, Trevor Field, who had acquired the patent to the invention, launched PlayPumps International and secured his first sponsor in 1995 and a World Bank grant in 2000. Initially, the PlayPump received widespread praise from international media, celebrities, and influential donors. High-profile supporters, including former U.S. President Bill Clinton and First Lady Laura Bush, brought global attention and millions of dollars in funding to Field's project. However, in 2009, reports documenting several shortcomings of the PlayPump were released (e.g., UNICEF, 2007, only made publicly available in 2009). The device was difficult and exhausting to use: children did not play on it as they found it tiresome or even unsafe, and women frequently ended up turning the pump themselves. Moreover, the PlayPump was considerably more expensive, less effective, and harder to maintain compared to traditional hand pumps. These reports eventually led to the US arm of PlayPumps International shutting down and its sponsor, the Case Foundation, withdrawing support. Yet, despite these setbacks, Field continued his mission under the new name Roundabout Water Solutions, still installing PlayPumps across South Africa with ongoing corporate support (for more details on the PlayPump story see MacAskill, 2015).

The adoption of the PlayPump raises many interesting psychological ques-

tions: Why did it take such a long time for people to recognize that the PlayPumps were an inferior alternative to existing pumps? Why did Field still continue installing the pumps even after the shortcomings were known? Why was initial enthusiasm for the project so high, even though at the time, the practical advantages of the PlayPump were arguably already relatively small compared to the substantial increase in cost (the PlayPump was around four times as expensive as conventional pumps)? What psychological interventions can help individuals make better decisions in similar situations going forward?

While this story can be seen as a cynical perspective on humans' moral and prosocial decision making, it actually has a relatively optimistic starting point: People take an issue faced by others in relatively distant countries with dissimilar cultures seriously and are motivated to act. This is and was, of course, not always the case for all types of moral decisions. For instance, consider slavery, which was widely accepted as a morally acceptable practice before British abolitionists successfully campaigned for it to be abolished in the British Empire at considerable economic cost (Blackburn, 2010, pp. 33–93; Fogel, 1994, p. 228). This raises another important question: How do people realise that practices which are commonly accepted at a certain time are morally wrong, and consequently, how is reliance on ethical principles (or decision mechanisms) adjusted?

Further, many of the underlying reasons for the failure to recognize the limitations of the PlayPumps program (e.g., difficulty of learning in situations without frequent accurate feedback, cf. Kahneman and Klein, 2009) and, arguably, also the wrongness of slavery, are not unique to limitations of moral and prosocial decision making. Instead, they can be seen as instantiations of more general decision making phenomena in the domain of moral and prosocial decision making. This PhD thesis adopts the view that usually the same principles apply to both non-moral, and moral and prosocial decision-making, and as a result, many chapters apply ideas and theories from broader decision-making research to moral and prosocial decisions.

My overarching goal in this thesis is to use a variety of tools available to a psychological or cognitive scientist (and in some parts those of an applied statistician)

to rigorously understand moral and prosocial decision making and interventions to increase prosociality. To do so, I will rely on two main approaches: publication bias-adjusted meta-analysis and experimental work combined with computational modelling. The first half of the thesis revisits areas where extant literature is already available (Construal Level Theory, the Identifiable Victim Effect, and nudging) and assesses the strength of evidence in these areas using methods that I developed to correct for publication bias. The second part of the thesis focuses on areas that have been relatively understudied, in particular, Unit Asking (a method to overcome scope insensitivity), decision making under extinction risk, and moral learning. In these areas, there is not enough past research for meta-analytic investigations (and often no consistent paradigm to study these topics). Instead, I will rely on a combination of experimentation and computational modelling.

0.1 Two Key Issues in Research on Moral and Prosocial Decision Making: Publication Bias and Lack of Formal Modelling.

This thesis is written with a particular focus on two issues in research on moral and prosocial decision making: publication bias and lack of specific theories implemented in computational models.

Publication bias, the selective publication of scientific results (usually in the form of preferentially publishing statistically significant results over non-significant results), leads to an overestimation of effect sizes when relying on the past literature. This is a particularly pervasive problem in meta-analysis, where aggregating the published literature will consequently lead to overestimation of evidence for and size of effects. Research has shown that publication bias, along with other problems such as questionable research practices, leads to substantial overestimation of the evidence for published effects in meta-analyses (Bartoš et al., 2023, 2024; Stanley et al., 2018). Therefore, it is pertinent that work relying on the past literature takes publication bias into account. To address this issue, my collaborators and I devel-

oped new methods to correct for publication bias in meta-analysis, which have been shown to outperform other methods in simulation studies and empirical examples (see Bartoš, Maier, Wagenmakers, et al., 2022; Maier, Bartoš, and Wagenmakers, 2023), and are applied in Part I of this thesis.

A second problem in research on moral and prosocial decision making is lack of **formal theory development**, in particular using **computational modelling**. In the history of psychology as a discipline, important breakthroughs have often come from the application of mathematical modelling (Navarro, 2021). This is unsurprising as implementing theories in computational models has several benefits: They allow making precise quantitative predictions, they allow comparing observed behaviour to optimality, and they allow researchers to infer parameters from data that may represent psychological constructs but cannot be easily inferred from data alone (Crockett, 2016; Tusche & Bas, 2021). Despite the general success of computational models in other areas of psychology to which they have been applied, there is relatively little work studying moral and prosocial decisions using these tools, with a few notable exceptions (see Awad et al., 2022; Crockett, 2016; Tusche & Bas, 2021). To close this gap, a key method in the second part of the thesis (in particular the chapters on moral learning and decisions under extinction risk) is reliance on computational models to understand behaviour on an individual level.

Importantly, publication bias and lack of formal modelling should not be viewed in isolation, but amplify each other in various ways. For instance, publication bias makes it more difficult to understand variability in effect sizes and consequently more difficult to develop formal theories to describe this variability. However, the lack of formal theory also amplifies publication bias as null effects become less informative if the theory is weak (in the General Discussion, I outline some ideas for an integrated methodological approach using publication bias adjustment and computational modelling).

0.2 Overview of the Thesis

This PhD thesis is structured in two parts. In the first part, I revisit important research areas in moral and prosocial decision making with publication bias adjustment methods. First, I reanalyse a meta-analysis on Construal Level Theory. This theory posits that psychological distance (associated, for instance, with distance in space or time) has various downstream consequences, such as increased abstraction (Jones et al., 2017; Liberman & Trope, 1998, 2014; Spence et al., 2012; Trope & Liberman, 2011), which influence decision making. It therefore has important implications for understanding how people think about spatially distant people (e.g., those living in faraway countries) and temporally distant people (e.g., those that will live in the future, Trope and Liberman, 2003, 2010).

Second, I revisit the Identifiable Victim Effect: the tendency to offer more support to an identifiable individual over a group of unidentified victims who are described using numerical statistics (Jenni & Loewenstein, 1997). This inconsistent valuation results in inefficient resource allocation in altruistic decisions. I reanalyse an important meta-analysis on this effect using methods to correct for publication bias and conduct a replication of a seminal paper.

In the third chapter of Part I, I investigate publication bias in research on nudging. Nudging describes behavioural interventions that aim to change behaviour without using incentives (for instance, by setting defaults). These nudges are promising tools for increasing prosocial behaviour (Capraro et al., 2019) and have been studied both in academic research and in government ‘nudge units’, such as the behavioural insights team and the Office for Evaluation Sciences (DellaVigna & Linos, 2022). In this chapter, I revisit both academic research on nudging and research by nudge units, finding strong publication bias in academic research and suggestive evidence for publication bias in research by nudge units.

Faced with the reality of publication bias and the strong impact it had on many fields in moral and prosocial decision making, I was faced with two possible options for how to continue in Part II: either I could revisit the same areas that I reanalysed in Part I with empirical studies to understand which types of effects are more likely

to replicate than others (to the extent that effects were heterogeneous), or I could create new paradigms studying other important areas of moral and prosocial decision making, which had been neglected in the previous literature. Ultimately, the thesis takes the second course for two main reasons. First, many of the most important topics related to moral and prosocial decision making have received too little attention in the previous literature (including the three areas that I study in the second part of my thesis). Second, for those areas that I studied using meta-analytic methods, adequate paradigms and methodologies were already in place to some extent. I was therefore more hopeful that other people would study these questions if I would not (and indeed there have been developments in that direction, such as a large multilab replication on Construal Level Theory, <https://climr.org/>), compared to the more neglected issues that I address in the second half of the thesis (which arguably would have remained neglected had I not started to tackle them here).

The second part, therefore, addresses three relatively understudied domain areas. First, I consider the problem of Scope Insensitivity, and in particular, the Unit Asking Method to address scope insensitivity. Scope Insensitivity has been put forward as one of the key reasons for many failures in altruistic decision making, from inefficient charitable donations (Västfjäll & Slovic, 2020) to our relative indifference in response to large-scale atrocities (Cameron & Payne, 2011; Dickert et al., 2012, 2015; Slovic & Västfjäll, 2010a).

Then, I investigate decisions under extinction risk. In the 21st century, humanity faces several risks that could, in the worst case, lead to human extinction, such as runaway climate change (Kemp et al., 2022), risks from artificial intelligence (Christian, 2021; Russell, 2019), bio-terrorism and pandemics (Millett & Snyder-Beattie, 2017a, 2017b), and nuclear war (Ellsberg, 2017; Perry & Collina, 2020). A recent survey of experts and super-forecasters¹ finds a median risk of human extinction by 2100 of 6% according to experts and 1% according to super-forecasters (Karger et al., 2023). These risks can be considered high for something as impactful as human extinction - most people would not take a flight with an airplane that

¹Superforecasters are individuals that are selected for excellent ability to forecast future events, but usually not domain experts (Tetlock & Gardner, 2016).

had a 6% or even a 1% chance of crashing. Therefore, this chapter aims to understand how individuals and groups make decisions under extinction risk and test interventions to reduce risk-taking in light of these risks.

In these two chapters, the ethically right decision can be seen as relatively straightforward (reduce extinction risk or donate to a charity that helps children in need); in the last chapter, I study scenarios, where this is not the case, and ask how people learn which decision mechanisms people rely on in these situations. In particular, I develop a new paradigm to study how people learn when to rely on moral rules (often equated with deontology) versus cost-benefit reasoning (often equated with utilitarianism). I further propose a theoretical model based on strategy selection learning (Lieder & Griffiths, 2017; Rieskamp & Otto, 2006), and implement it in reinforcement learning models to understand individual differences in learning styles and how they affect generalisation.

All chapters of this dissertation are based on collaborative work (though I was first or shared-first author on all published projects included here). I will use the pronoun ‘we’ in the chapters describing work that I conducted together with collaborators, while I use ‘I’ in the introduction and discussion. Author contributions are stated at the end of each chapter. In addition to the Appendices, many sections include online supplementary information. The link to the supplementary information is always provided at the end of the corresponding section.

Part I

Revisiting Classic Theories and Interventions with Novel Publication Bias Adjustment Methods

Background for Part 1

As outlined in the general introduction, I consider publication bias one of the key challenges for research in moral and prosocial decision making. This chapter, therefore, applies publication bias correction methods to adjust for this bias in various research areas related to moral and prosocial decision making.

For most of the reanalyses, I rely on a method called Robust Bayesian Meta-Analysis (RoBMA), which my collaborators and I developed. RoBMA uses Bayesian model-averaging to combine Bayesian implementations of multiple frequentist publication bias adjustment methods. The key idea of Bayesian model-averaging is applying an ensemble of models to the data simultaneously (36 models in RoBMA), whereby the weight given to each model is based on its posterior probability, which is proportional to the marginal likelihood (i.e., the probability of the data given the model). This approach allows the data to guide the inference to be based most strongly on those models that predicted the data best and makes RoBMA uniquely robust to model misspecification.

RoBMA has been extensively tested in simulation studies as well as on empirical data by comparing publication bias-adjusted meta-analysis to registered reports. In all these demonstrations, RoBMA has outperformed other methods across most meta-analytic settings (Bartoš, Maier, Wagenmakers, et al., 2022; Maier, Bartoš, & Wagenmakers, 2023). Nevertheless, publication bias-adjusted meta-analysis is no panacea, and ultimately, it is important to conduct direct replications and adopt methods to mitigate publication bias, such as registered reports. I therefore do not consider the demonstrations in this part conclusive evidence but rather cautionary notes that should motivate further research in the form of direct replications. For the Identifiable Victim Effect, we follow the RoBMA reanalysis with a replication, while for the other topics, replication efforts, which were in part motivated by the meta-analytic work presented here, are currently ongoing (see <https://climr.org/>).

Including all the statistical papers on development of the RoBMA method would be beyond the scope of the current thesis, but I include the references to key papers below, for anyone who wants to consult them alongside the demonstrations

in Part I:

Maier, M.*, Bartoš, F.*, & Wagenmakers, E.-J. (2023). Robust Bayesian Meta-Analysis: Addressing publication bias with model-averaging. *Psychological Methods*, 28(1), 107-122. <https://doi.org/10.1037/met0000405>.

Bartoš, F., **Maier, M.**, & Wagenmakers, E.-J.(2023). Robust Bayesian Meta-Analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, 14, 99-116 <https://doi-org.libproxy.ucl.ac.uk/10.1002/jrsm.1594>.

Bartoš, F.*, **Maier, M.***, Wagenmakers, E.-J., Doucouliagos, H., Stanley, T.D. (2022). Adjusting for publication bias in JASP and R: Selection models and Robust Bayesian Meta-Analysis. *Advances in Methods and Practices in Psychological Science* 5(3), 1-19. <https://doi.org/10.31234/osf.io/75bqn>.

Chapter 1

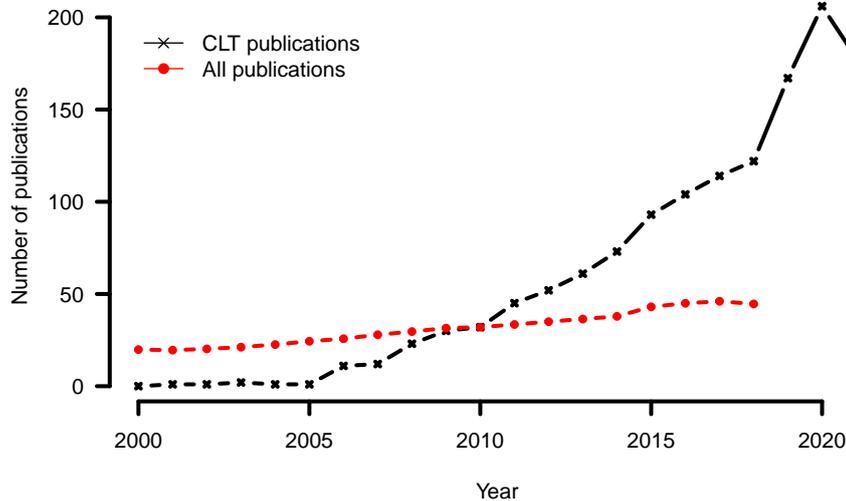
Construal Level Theory

Construal Level Theory (CLT) is one of the most prominent social psychology and decision making theories. In recent years, a large number of studies have been inspired by this theory (Figure 1.1), and the foundational review article (Trope & Liberman, 2010) has been cited > 8000 times at the time of writing (Google Scholar, May 19th, 2025).

CLT proposes that thinking about temporal distance, spatial distance, another person's perspective (interpersonal distance), and counterfactual alternatives all constitute different forms of traversing psychological distance and as such invoke the same underlying psychological mechanisms. Greater psychological distance leads to a higher construal level or more abstract representation of phenomena (Liberman & Trope, 1998, 2014; Trope & Liberman, 2003, 2010, 2011). For example, participants indicated a greater preference to describe the activity, 'filling out a personality test' as 'revealing what you're like' (rather than 'answering questions') when thinking about their life further in the future (Sánchez et al., 2021). This straightforward theoretical framework has inspired a variety of applications of psychological science to real-world problems, including important issues such as the way people think about infectious diseases (e.g., Cai & Leung, 2020; Z. Li et al., 2020; Van Lent et al., 2017) and climate change (e.g., Brügger, 2020; Brügger et al., 2016; Chu & Yang, 2019; Jones et al., 2017; Keller et al., 2021; Loy & Spence, 2020; Maiella et al., 2020; Reczek et al., 2018; Rickard et al., 2016; Schuldt et al., 2018; Spence et al., 2012; Wang et al., 2019), with some even stating that 'the

key problem [underlying climate change] is psychological distance' (Van Lange & Huckelba, 2021, p.49).

Figure 1.1: Research on Construal Level Theory Is Increasing Rapidly



Note. Increase of publications on Construal Level Theory (black) in comparison to the general increase in scientific output (red) based on Bornmann et al. (2021) and scaled to overlap in 2010.

1.1 Motivation for the Current Research

In light of the theoretical and applied importance of CLT, it is crucial to rigorously assess the evidence underlying this theory. So far, only a few replication efforts have been directed towards studies supporting CLT and they show a mixed pattern.

Some replication studies have directly examined the effect of psychological distance on construal level. Early small studies consistently found that increased psychological distance led to higher levels of abstraction (Liberman & Trope, 1998; Liviatan et al., 2008; Wakslak et al., 2006). In addition, a high powered replication showed a moderate effect of $d = 0.276$ (Sánchez et al., 2021) of temporal distance. However, another higher-powered replication did not observe the predicted effect of likelihood on abstraction (Calderon et al., 2020). A recent article suggests that the effect of likelihood on abstraction might occur only when the study explicates the contrast between proximal and distal by including both high and low likelihood for reference (Grinfeld et al., 2021).

Two further replications of CLT studies have been conducted in the domain of

moral psychology. T. Eyal et al. (2008) showed that virtuous acts are judged more positively, and moral transgressions more negatively, when psychological distance is large rather than small. This effect is predicted by CLT on the basis that people tend to construe distant acts more in terms of moral principles rather than considering situation-specific factors (T. Eyal et al., 2008). However, Gong and Medin (2012) found the opposite result. Žeželj and Jokić (2014) conducted a third replication that tried to reconcile these findings. They found yet a different pattern of results: no effect of temporal distance on moral judgement, but an effect of social distance in line with CLT predictions, and the opposite effect for a direct manipulation of construal level. The inconsistent and partially opposing findings across a set of three studies prompted a series of commentaries (T. Eyal et al., 2014; Gong et al., 2014) proposing different mechanisms to recast these findings (a detailed examination of which is beyond the scope of this short summary).

In the domain of climate change communication, Yang et al. (2021) failed to replicate the effects of psychological distance on mental construal using a cued distance manipulation originally devised by Schuldt et al. (2018). Further, a recent review finds that the results of applications of CLT to climate change are mixed and contradictory (Maiella et al., 2020).

Next to these direct replications, conceptual replications provide additional evidence. Notably Snefjella and Kuperman (2015) find support for CLT predictions in different domains in a large study analyzing natural language use. However, while mostly in line with CLT predictions, their study finds a curvilinear relationship with a small increase of concreteness for large distances. In addition, Bhatia and Walasek (2016) find strong support for a relationship between temporal distance and abstraction using large natural language corpora. While these studies are an important piece of evidence due to their large sample sizes (all studies $N_s > 1\,000\,000$), they suffer from the drawback of being conceptual rather than direct replications. If a conceptual replication fails, it is usually less informative than a direct replication as it is unclear whether this reflects failure to replicate the underlying theory or problems with the idiosyncratic adaptations made in this replication. This

asymmetric informativeness may cause favorable publishing of conceptual replications that are successful (Chambers, 2019; Pashler & Harris, 2012).

This section shows that evidence from high-powered replications is only available in a few domains and in these areas the patterns of evidence are often conflicting. In addition, the wide set of predictions and implications that can be derived from CLT make it difficult to assess credibility using large-scale replications alone. However, given the substantial influence of CLT, it is important to obtain a wider overview of its evidential standing. Therefore, this chapter takes a different approach. Here we ask whether the available literature on CLT shows any evidence of being tainted by publication bias. To the extent that publication bias exists in a field, with positive findings being selectively published relative to null or negative findings (the file-drawer effect, R. Rosenthal and Gaito, 1964), the published research will present an inflated estimate of the robustness of the key effects. Hence the present work adds to the existing evidence base by examining the evidence for underlying effects in CLT studies with novel publication bias detection methods.

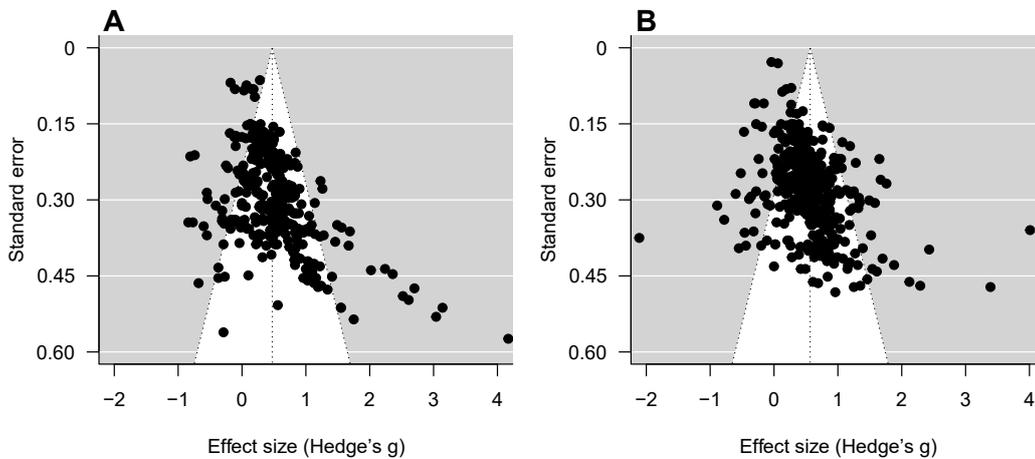
1.2 Reanalysing a Comprehensive Meta-Analysis on Construal Level Theory

Soderberg et al. (2015, cited more than 380 times at the time of writing, May 17th, 2025) represents the most recent and largest meta-analysis of the effects of psychological distance on abstraction (310 estimates) and the downstream consequences of abstraction (426 estimates). Consequently, Soderberg et al. (2015) provides an appropriate dataset for an initial exploration of the potential role of publication bias in CLT research. From their meta-analysis, Soderberg et al. (2015, p. 525) reported ‘a reliable and medium-sized effect of psychological distance on both level of abstraction in mental representation and the downstream consequences of abstraction’. But is this conclusion robust to the possible presence of publication bias?

Soderberg et al. (2015) explicitly address the potential for publication bias in their meta-analysis with methods available at the time. They provide a funnel plot (Figure 1.2), demonstrating a considerable skew in effect sizes, which is indicative

of potential publication bias. Soderberg et al. (2015) subsequently quantified the influence of the suggested publication bias via the trim-and-fill method (Duval & Tweedie, 2000), finding that there were only 39 ‘missing’ studies whose inclusion would reduce the meta-analytic effect size to $g = 0.30$. In addition, using fail-safe N (R. Rosenthal, 1979) they showed that 15,202 studies with a null effect size would have needed to be missing to reduce the aggregate effect to zero.

Figure 1.2: Funnel Plots for Soderberg et al. (2015)



Note. A: The influence of (A) construal level on abstraction. (B) Downstream consequences of abstraction. Five (A) and eight (B) estimates are not shown as they were outside the plotting range.

There is, however, a large and rich literature on bias correction methods which shows that both of these approaches, as well as visual inspection of funnel plots, are insufficient to detect and correct for publication bias and often have high false positive rates (concluding that a true effect exists when it does not; e.g., Bartoš, Maier, Quintana, and Wagenmakers, 2022; Bartoš, Maier, Wagenmakers, et al., 2022; Carter et al., 2019; Hong and Reed, 2020; Hunter and Schmidt, 2004; Kvarven et al., 2020; Lau et al., 2006; Maier, Bartoš, and Wagenmakers, 2023; Mathur and VanderWeele, 2020).

While a number of alternative approaches have been suggested in the literature (for a review, see Carter et al., 2019), a shared problem for these approaches is that they only perform well in some meta-analytic conditions (i.e., type of publication bias and heterogeneity). However, it is not possible to know the meta-analytic conditions without having adjusted for publication bias, which poses a conundrum to

model selection (Bartoš et al., 2023). The recent Robust Bayesian Meta-Analysis (RoBMA) approach (Bartoš, Maier, Quintana, & Wagenmakers, 2022; Bartoš, Maier, Wagenmakers, et al., 2022; Bartoš & Maximilian, 2020; Maier, Bartoš, & Wagenmakers, 2023) is an advance on previous methods because it uses Bayesian model-averaging to apply multiple methods at the same time and is therefore the approach employed here. RoBMA combines selection models (models that use a weighted likelihood to account for studies that are missing due to publication bias; Iyengar & Greenhouse, 1988; Maier et al., 2022; Vevea & Hedges, 1995) and PET-PEESE (a method that employs linear and quadratic regression to correct for publication bias based on the relationship between effect sizes and standard errors; Stanley, 2017; Stanley & Doucouliagos, 2014) under the Bayesian model-averaging umbrella (Hinne et al., 2020; Hoeting et al., 1999). Rather than selecting a single method, Bayesian model-averaging combines the estimates obtained from different methods based on their posterior probability. In other words, models contribute to the inference in proportion to how likely they are to have generated the data. This allows the data to guide the inference to be based most strongly on those models that capture the publication process best. RoBMA allows us to draw inferences about the presence versus absence of the effect, heterogeneity, and publication bias.

RoBMA's ability to efficiently correct for publication bias has been demonstrated repeatedly in simulation studies as well as applied examples. First, in simulations, Bartoš, Maier, Wagenmakers, et al. (2022) tested RoBMA by reproducing a large simulation by Hong and Reed (2020), which combined four previous simulation studies (Alinaghi & Reed, 2018; Bom & Rachinger, 2019; Carter et al., 2019; Stanley et al., 2017). This indicated that RoBMA performs well under a wide range of settings and outperforms other bias correction methods most of the time in terms of bias and root mean squared error. RoBMA's performance was also evaluated on empirical data. Kvarven et al. (2020) compared the accuracy of meta-analytic estimates to the estimates of registered replication reports (RRR, Chambers, 2013; Chambers et al., 2015) – often considered the gold standard of empirical evidence. Bartoš, Maier, Wagenmakers, et al. (2022) reproduced this analysis with RoBMA,

showing that RoBMA’s meta-analytical estimate tracks the RRRs substantially better than naive random-effects meta-analysis and better than other bias correction techniques. Finally, Maier, Bartoš, and Wagenmakers (2023) tested the false positive rate of different bias correction techniques using studies from Many Labs 2 (Klein et al., 2018). Many Labs 2 is a collection of registered replication reports; therefore, we know that publication bias is absent. If a bias correction technique finds evidence for bias nevertheless, this is likely to be a false positive. This analysis showed that whereas many other methods detect publication bias where none is possible, RoBMA maintained a low false positive rate.¹

When RoBMA is applied to the Soderberg et al. (2015) dataset about the effects of psychological distance on abstraction, the results indicate: (1) very strong evidence for the presence of publication bias, $BF_{pb} > 10^{100}$,²³ (2) moderate evidence for the absence of the effect, $BF_{01} = 7.75$; and (3) a model-averaged effect size of $\delta = -0.002$, 95% $[-0.113, 0.056]$. In other words, the data are 7.75 times more likely under the null hypothesis of no effect than under the alternative prior of $\text{Normal}(0,1)$, and the estimated true effect is close to zero.⁴

When considering the downstream consequences of abstraction, RoBMA also revealed overwhelming evidence for publication bias, $BF_{pb} > 10^{100}$, but with evidence in favor of the effect for downstream consequences of abstraction, $BF_{10} = 16.29$; however, the resulting effect size estimate was in the direction *opposite* that in the original meta-analysis, $\delta = -0.360$, 95% $[-0.657, 0.000]$. When a bias cor-

¹RoBMA might be criticised for being too conservative (i.e., correcting effect sizes too strongly). However, evidence from the comparison of RRRs to bias-corrected meta-analyses as well as the low false positive rate of the test for publication bias on RRRs shows that this is not the case. In addition, RoBMA often finds no bias in published meta-analyses. For example, Shipley and van Riper (2022) show that pride and guilt predict pro-environmental behaviour; a reanalysis with RoBMA comes to the same conclusion for both experimental and correlational studies with no evidence of bias in either of the domains or study types.

²The publication bias correction models were preferred beyond the numerical precision of \mathbb{R} (around 10^{308}).

³As a general rule of thumb, Bayes factors between 1 and 3 are regarded as anecdotal evidence, Bayes factors between 3 and 10 are regarded as moderate evidence and Bayes factors larger than 10 are regarded as strong evidence. When evidence for the null is considered the Bayes factor is inverted (i.e., $BF_{01} = 1/BF_{10}$) (e.g., Jeffreys (1939); M. D. Lee and Wagenmakers (2013, p. 105); Wasserman (2000)).

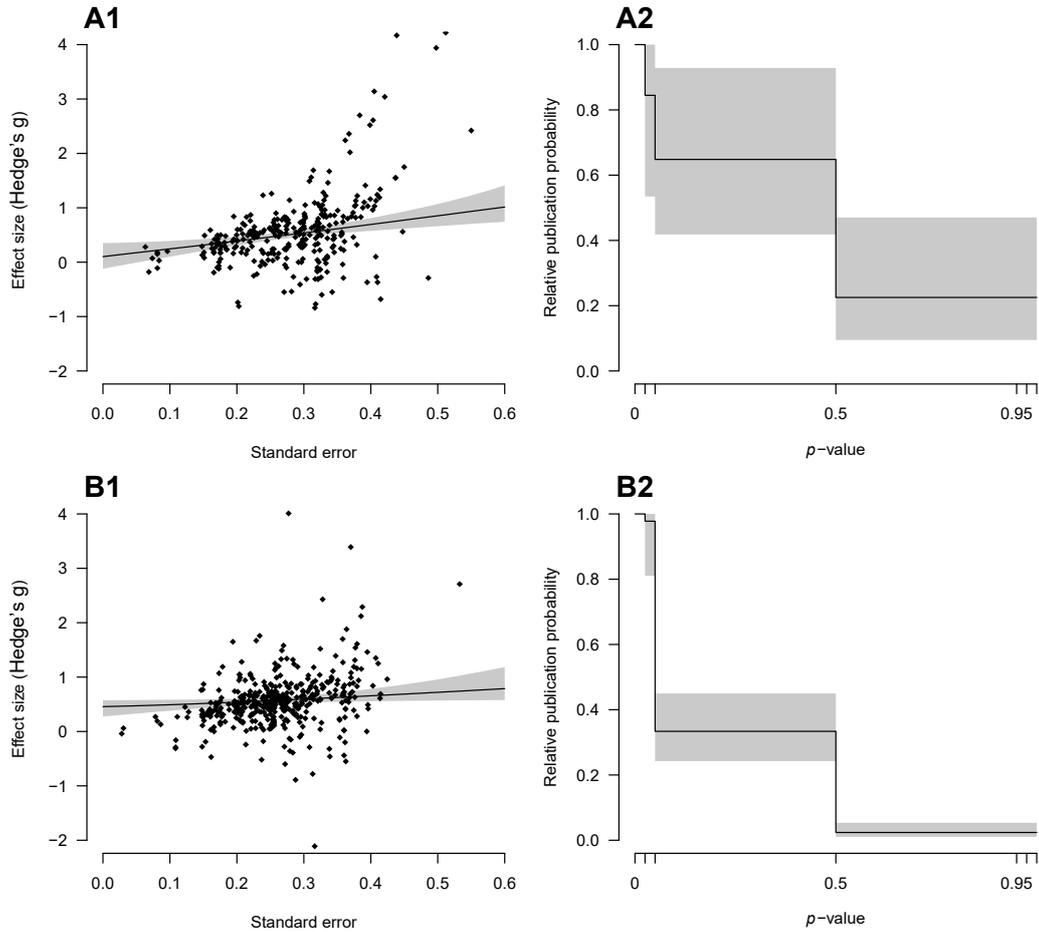
⁴For more information about prior specification, see Bartoš, Maier, Wagenmakers, et al. (2022) Appendix B.

rection technique reverses the estimate of an effect size to the opposite direction, this should usually be considered as indicative of a null effect (e.g., Carter et al., 2019).⁵

Figure 1.3 shows the footprints of publication bias in the Soderberg et al. (2015) dataset. Panel 1 visualises the RoBMA model-based estimate of the relationship between standard errors and effect sizes; the results show that studies with smaller sample sizes (and more noisy estimates) systematically lead to higher (biased) effect size estimates. In addition, the intercept of the PET-PEESE regression line shows that the hypothetical effect size for a study with infinite precision would be tiny for the effects on abstraction (A1). Panel 2 visualises the relationship between one-sided p -values and relative publication probabilities. This visualization suggests that studies with statistically non-significant effect size estimates are considerably less likely to be published than significant ones, especially in the studies on downstream consequences (B2). This pattern is especially severe for studies with effects in the opposite direction to what CLT predicts (as signified by the fact that studies with p -values larger .5, which indicates effects opposite the predicted direction, have very low relative publication probabilities).

⁵This correction in the opposite direction is likely due to the extreme bias observed based on p -value selection. The relative publication probability of studies estimating effects in the opposite direction is estimated to be extremely low; therefore, adjusting for this difference in publication probability flips the overall estimate to be in the opposite direction. As Table A.2 indicates, the estimate is corrected to zero, when all effects $|d| > 1.5$ are removed, indicating that the interpretation as a true zero effect is more appropriate than concluding an effect in the opposite direction.

Figure 1.3: Footprints of Publication Bias in the Soderberg et al. (2015)



Note. Column 1 shows the relationship between standard errors (x -axis) and effect sizes (y -axis) based on the conditional RoBMA PET-PEESE models (five and eight estimates out of the plotting range not shown); Column 2 shows the relationship between one-sided p -values (x -axis) and relative publication probabilities (y -axis) based on the conditional RoBMA selection models. Row A shows the results for the effect of psychological distance on abstraction, and row B shows the results for the downstream consequences of abstraction.

Notably, both meta-analyses provided extreme evidence in favor of heterogeneity with the between study heterogeneity estimate $\tau = 0.367$, 95% [0.314, 0.421], and $\tau = 0.712$ [0.616, 0.804], for the effects on abstraction and downstream consequences, respectively. Therefore, we also conduct subgroup analyses based on the moderators identified in Soderberg et al. (2015), specifically: Setting (lab, field, or online), ‘Real or Imagined’ (i.e., whether the manipulation of distance in the study was perceived by the participants to be real or imaginary), and ‘Focus of Measure’ (i.e., whether the dependent variable measured changes in high level characteristics, low level characteristics or both). The subgroup analysis

allows us to test whether the mean effect in any of the groups is different from zero and allows for the publication process to operate differently depending on the levels of the moderator. Table 1 shows the results of this analysis. The only group that shows evidence for a mean effect larger than zero is ‘field studies’. Further, we find evidence for heterogeneity ($BF_{10} > 3$) in 8 out of 11 subgroups.⁶

Due to the inability of publication bias adjustment methods to perform a 3-level meta-analysis that accounts for clustering of effect size estimates within individual studies (the previous results ignored the dependencies between individual effects), we also re-analysed the data in two additional ways: (1) selecting the most precise estimate from each study; (2) aggregating the estimates from each study (using their precision weighted average, following Soderberg et al.). In addition, Soderberg’s meta-analysis contained several unrealistically large effects which might have a strong impact on the PET-PEESE regression line. Therefore, we also analysed the data with all effects larger than $|d| = 1.5$ removed (19 estimates for the effect of psychological distance on abstraction and 21 estimates for downstream consequences of abstraction). Finally, because regression-based methods are sometimes criticized for finding bias for reasons other than publication bias (Lau et al., 2006), we also reanalysed the data by only using the selection models in RoBMA (excluding PET-PEESE). Overall, this resulted in 12 different analyses (2x2x3). The results of this ‘RoBMA multiverse’ (Gelman & Loken, 2013; Orben & Przybylski, 2019; Wagenmakers et al., 2022) can be found in Appendix A.1. All of these permutations show very strong evidence for publication bias and smaller effect sizes than the original meta-analysis. Regarding the effect of psychological distance on abstraction, six analyses show evidence against an effect,⁷ three are undecided and three show evidence in favor. For downstream consequences of abstraction, six show evidence against an effect, two are undecided, and four show evidence in favor of an effect. In other words, the CLT literature is certainly contaminated by publication bias, which leads to strong inflation of effect sizes. However, while the default analysis

⁶The justifications for focusing only on the significant moderators are (a) that we would not expect a non-significant moderator to show evidence after adjusting for publication bias, and (b) that including all moderators and sensitivity analyses for them would require more than 200 analyses.

⁷Including two cases that show evidence for an effect in the opposite direction.

and the majority of additional analyses show evidence of absence regarding a mean effect, this finding is not entirely robust to model specification. Specifically, when removing effect sizes larger than $\delta = 1.5$, we find moderate to strong evidence for an effect for some of the model specifications (see Appendix A.1 for details). Overall, the reanalysis of Soderberg et al. (2015) thus shows that publication bias leads to considerable overestimation of CLT effect sizes, and may even reduce mean effects to zero.

1.3 Sequential Robust Bayesian Meta-Analysis of Recent Studies

Given the recent credibility revolution and improvement of the quality of evidence in behavioural research (Nosek et al., 2018, 2022; Vazire, 2018), it is crucial to examine more recent studies than those included in Soderberg et al. (2015). To assess more recent evidence for Construal Level Theory, we conducted a sequential meta-analysis. In frequentist meta-analysis, sequential updating and testing as the data accumulate inflates the rate of false positives; however, evaluating evidence using the Bayes factor permits sequential analysis to be undertaken (Rouder, 2014; Wagenmakers et al., 2016).

We used the same search terms as in Soderberg et al. (2015) and coded studies backwards in time starting in July, 2022. We terminated the data collection after collecting 41 downstream consequences estimates and 10 abstraction estimates. We initially aimed to code until we reached a Bayes factor of 3 for either the null or the alternative hypothesis of the presence of the effect under the Osterwijk prior. However, studies on the effect of construal level on abstraction have become more rare in the recent literature (likely because researchers see this effect as established and focus on studying novel downstream effects). Therefore, we only found ten estimates for the effect of construal level on abstraction even when going back until June, 2019.⁸

⁸We coded both effects on abstraction and downstream consequences of abstraction in articles published from July, 2022 until August, 2021 after we switched to looking only at effects of construal level on abstraction, given the scarcity of studies on this effect after June, 2019. We stopped the cod-

To increase our ability to find evidence for the null or the alternative model of the effect, and thus the efficiency of the procedure, we use the informed Oosterwijk prior distribution on the standardized mean difference (Gronau, van Erp, et al., 2017). The Oosterwijk prior distribution, Student- $t(\mu = 0.350, \sigma = 0.102, \nu = 3)$, was specifically elicited from a social psychologist to describe small to medium effects in social psychology. To ascertain the robustness of our findings, we also re-analysed the data with the default RoBMA settings, a Normal($\mu = 0, \sigma = 1$) prior distribution on the effect. The conclusions are the same in terms of the Bayes factor categories unless stated otherwise in a footnote.

The recent studies on downstream consequences showed moderate evidence against the presence of the effect, $BF_{01} = 6.71$,⁹ with a mean model-averaged estimate of Cohen's $d = 0.032$, 95% CI [0.000, 0.342]. Furthermore, we found extreme evidence in favor of heterogeneity, $BF_{rf} = 7.51 \times 10^{93}$, with a mean model-averaged estimate of $\tau = 0.62$, 95% CI [0.486, 0.778]. Finally, we found strong evidence in favor of publication bias, $BF_{pb} = 19.56$.

The recent studies on the effect of construal level on abstraction showed weak evidence for the presence of the effect, $BF_{10} = 1.347$,¹⁰ with a mean model-averaged estimate of Cohen's $d = 0.148$, 95% CI [0.000, 0.400]. Furthermore, we found extreme evidence in favor of heterogeneity, $BF_{rf} = 2187.11$, with a mean model-averaged estimate $\tau = 0.252$, 95% CI [0.121, 0.457]. Finally, we found moderate evidence in favor of publication bias, $BF_{pb} = 3.61$.¹¹

ing at this point as we considered earlier studies to be less likely impacted by recent improvements in research practices.

⁹Weak evidence against the presence of the effect, $BF_{01} = 2.71$, when using the default RoBMA settings.

¹⁰A weak evidence against the presence of the effect, $BF_{01} = 2.27$, when using the default RoBMA settings

¹¹A strong evidence for the presence of publication bias, $BF_{01} = 11.83$, when using the default RoBMA settings

1.4 Quantitative Assessment of Publication in a Broader Set of Studies using Z-Curve

Our reanalysis of the dataset by Soderberg et al. (2015) focuses only on the effect of abstraction and the downstream consequences. In addition, to obtain a sense of the complete research field, it is also important to examine a wider range of studies, with all interventions and dependent measures that could be associated to CLT.

To investigate more recent evidence, and also to obtain a wider overview of the CLT literature, we coded the 400 most highly-cited studies (in terms of citations/year) on CLT published since 2013 for a z-curve analysis. We choose to focus on the most highly-cited studies, rather than for example a random subset, as we thought it most important to check the robustness of those studies that researchers will read and cite in practice.

Z-curve estimates the mean statistical power of the published studies after selection for statistical significance (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020). Unlike standard meta-analytic methods such as RoBMA, z-curve does not assume the studies originate from a single distribution of effects. Therefore, it is uniquely suited for assessing the credibility of literature in contexts where different studies may test completely different effects. This is not the case of RoBMA which assumes that the individual studies are assessing a common underlying construct. Therefore, we conducted the following wider assessment of the CLT literature with z-curve.

To estimate the mean statistical power of a diverse set of studies and account for a possibly large degree of heterogeneity, z-curve models the distribution of test statistics. First, z-curve transforms the test-statistics into z-scores and then models them as a mixture of truncated folded normal distributions. In other words, z-curve's goal is to approximate the overall distribution of all statistically significant studies regardless of the direction of the effect (under the assumption that all significant studies get published). This approximate distribution is then used to compute summaries about the published statistically significant studies, the expected replication

rate, and to extrapolate and provide information about all conducted (and possibly unpublished) studies, the expected discovery rate.

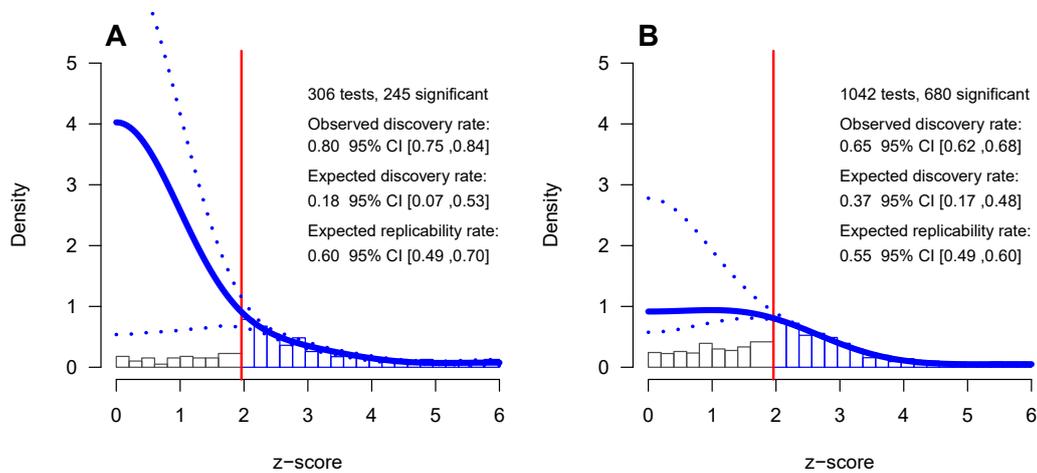
Z-curve quantifies the degree of publication bias in terms of the expected replication rate and expected discovery rate. The expected replication rate is the proportion of statistically significant studies that are expected to produce a statistically significant result in an exact replication with the same sample size. In other words, it corresponds to the unconditional power of statistically significant studies (i.e., power not assuming \mathcal{H}_1 to be true). The expected discovery rate reflects the expected proportion of statistically significant studies amongst all conducted (and possibly unobserved) studies. A lower expected discovery rate than the observed discovery rate indicates that a proportion of non-significant studies is likely to be missing from the literature (Bartoš & Schimmack, 2022).

The accuracy of the expected replication rate estimates under heterogeneity was corroborated in a simulation study by Brunner and Schimmack (2020). Later, Bartoš and Schimmack (2022) extended these simulations and tested z-curve's performance in both the expected discovery rate as well as the expected replication rate in cases with studies generated from true null hypotheses. Furthermore, Bartoš and Schimmack (2022) used two large data sets of hand-coded test statistics from social psychology journals and found that the expected replication rate estimated by z-curve overestimated the observed replication rate of social psychology studies from the Open Science Collaboration project, which suggests that, if anything, inferences from z-curve may yield conservative estimates of the degree of publication bias in a literature (Open Science Collaboration, 2015).

We obtained 306 test statistics of focal-hypotheses and conducted a z-curve analysis to estimate replicability of the selected studies from the published literature (see Appendix A.3). To achieve a more powerful assessment of the estimated replicability across an extended period, we further combined the newly coded test statistics with data from the Soderberg et al. (2015) meta-analysis. Figure 1.4 compares histograms of the observed distribution of test-statistics (converted to z -scores) to the z-curve based estimate of the expected distribution of test-statistics (full blue

line). Results of both z-curve analyses suggest considerable publication bias indicated by the mismatch between the expected discovery rate and the observed discovery rate; 18%, 95% CI [7%, 53%], vs. 80%, 95% CI [75%, 84%], based on the newly recoded studies (Panel A), and 37%, 95% CI [17%, 48%], vs. 65%, 95% CI [62%, 68%], based on the extended data set (Panel B). However, the estimated replication rate of 60%, 95% CI [49%, 70%] based on the newly recoded studies (Panel A), and 55%, 95% CI [49%, 60%], based on the combined data set (Panel B), shows that some studies on Construal Level Theory might hold up under replication.

Figure 1.4: Z-Curve Analysis of Test-Statistics from Construal Level Theory Studies



Note. Estimated distribution (and 95% CI) of test statistics based on statistically significant results (blue line) vs. observed distribution (histogram). The much lower number of observed than expected nonsignificant z-scores indicates the existence of studies missing due to publication bias. Panel A shows z-curve fitted to the 306 hand-coded test statistics of focal hypothesis tests from the 400 most highly-cited articles on Construal Level Theory published since 2013. Panel B shows z-curve fitted to an expanded data set including these 306 effects as well as test statistics from Soderberg and colleague's (2015) meta-analysis.

1.5 Conclusion

We critically assessed the empirical evidence underlying Construal Level Theory. Our first conclusion is that the evidential value of studies in this field is substantially reduced after adjusting for publication bias. First, a RoBMA multiverse re-analysis of a large dataset – previously interpreted as providing strong evidence for CLT – found strong evidence of publication bias. In addition, most model specifications show evidence against both an effect of construal level on abstraction and on down-

stream consequences of construal level after publication bias is accounted for. For those model specifications where we did not find evidence of absence, the effect size was still considerably smaller than in the original meta-analysis. For the downstream effects this finding is further corroborated by a sequential meta-analysis of more recent studies, which shows evidence against an overall effect on downstream consequences. Further, we quantitatively assessed more recent evidence with a z-curve analysis. This analysis indicated a striking mismatch between the expected and observed discovery rates – a pattern that is indicative of publication bias.

Importantly, it is not clear which part of the research process causes the estimated low replicability and discovery rates. It is possible that CLT as an idea is fundamentally correct, but that poor measurement or weak manipulations cause low expected discovery rates in practice (Eronen & Bringmann, 2021; Flake & Fried, 2020; Lilienfeld & Strother, 2020). For example, measures of concreteness have been criticized for lack of validity (Yeomans, 2021). In addition, many CLT experiments employ between-subjects priming procedures (e.g., imagine oneself engaging in and describing activities either ‘tomorrow’ or ‘next year’ as the manipulation Liberman and Trope, 1998 or viewing a map that is displayed on a larger vs smaller screen Schuldt et al., 2018), and recent large scale replications show that between-subjects priming effects are often not replicable (Camerer et al., 2018).

In addition, the extremely large heterogeneity in the Soderberg et al. (2015) meta-analysis questions whether we can draw strong conclusions about individual studies based on evidence regarding a mean effect. In other words, due to the large variability of true effects, some CLT methods are likely still effective (while others probably show backfire effects) even if the mean effect was zero. We find that heterogeneity persists for most subgroups in a subgroup analysis. Better understanding the heterogeneity of CLT effects is therefore a key challenge for future research. To this end, we are currently developing moderator analyses and mixture modelling for RoBMA.

Finally, bias correction techniques are no panacea. Importantly, the estimates from these methods cannot be seen as an alternative to high-powered replications

(Carter et al., 2019; Kvarven et al., 2020). Therefore, it is important to carry out more empirical research to critically assess the evidence underlying CLT. This new research should ideally be conducted via high-powered registered reports to arrive at a more credible assessment of the theory (Chambers, 2013; Chambers et al., 2015). We especially welcome one such initiative (<https://climr.org/>), which examines the effect of the four different distance domains on abstraction and plans to finish data collection in late 2023. We hope that more replication efforts like this will also be initiated on the applied implications of the theory and downstream consequences of abstraction. Until these replication projects are conducted, researchers should treat CLT based interventions with caution, given the doubts raised here about the underlying evidence base.

Reference to Preprint: Maier, M., Bartoš, F., Oh, M., Wagenmakers, E., Shanks, D., & Harris, A. J. L. (2022). Adjusting for Publication Bias Reveals That Evidence for and Size of Construal Level Theory Effects is Substantially Overestimated. <https://doi.org/10.31234/osf.io/r8nyu> [unpublished preprint]

Data and Code Availability: Data and analysis code are available at <https://osf.io/sd4yc/>.

Author contributions: M.M. designed research and conceptualized the project; M.M. and F.B. performed research; M.M. and F.B. analysed data; and M.M., F.B., T.D.S., D.R.S., A.J.L.H., and E.-J.W. wrote the paper.

Table 1.1: Subgroup Analysis based on Significant Moderators in Soderberg et al. (2015)

Abstraction	Moderator	δ (95% CI)	τ (95% CI)	BF ₀₁	BF _{rf}	BF _{pb}
Overall		-0.00 [-0.11, 0.06]	0.37 [0.31, 0.42]	7.78	$> 10^{100}$	$> 10^{100}$
Setting	Lab	0.01 [-0.05, 0.17]	0.35 [0.29, 0.41]	8.24	3.06325×10^{17}	1.39×10^6
	Field	0.39 [0.00, 0.77]	0.23 [0.00, 0.73]	0.29	2.916	1.437
	Online	-0.00 [-0.10, 0.09]	0.30 [0.13, 0.48]	7.66	1.28×10^4	3.556
Real or Imagined	Imagined	-0.00 [-0.16, 0.18]	0.46 [0.39, 0.53]	5.58	1.03×10^4	2.59×10^5
	Real	0.10 [0.00, 0.29]	0.04 [0.00, 0.14]	0.71	0.959	63.528
Focus of Measure	High Level	0.00 [-0.05, 0.10]	0.07 [0.00, 0.22]	9.17	1.548	311.904
	Low Level	-0.00 [-0.18, 0.12]	0.16 [0.00, 0.29]	5.00	8.613	2.21×10^5
	Relative	0.05 [0.00, 0.34]	0.56 [0.42, 0.67]	2.47	1.11×10^{40}	7104.547
Downstream Consequences						
Overall		-0.36 [-0.66, 0.00]	0.71 [0.62, 0.80]	0.06	10^{100}	10^{100}
Setting	Lab	-0.05 [-0.41, 0.00]	0.56 [0.49, 0.67]	3.41	1.61×10^{61}	1.46×10^{17}
	Field	-0.08 [-0.78, 0.17]	1.03 [0.84, 1.28]	2.41	2.48×10^{168}	3168.637
	Online	-0.00 [-0.08, 0.00]	0.10 [0.00, 0.22]	11.07	3.175	1.77×10^7

Chapter 2

Identifiable Victim Effect

In the previous chapter, we demonstrated how publication bias affects a key theory, which explains how people make judgments about distant others. This chapter builds on that insight by examining a possible strategy for reducing psychological distance – making victims more identifiable. Specifically, we explore how publication bias may also shape findings in the literature on the Identifiable Victim Effect and replicate a seminal paper on this effect.

The Identifiable Victim Effect is the tendency to offer more support to an identifiable individual over a group of unidentified victims who are described using numerical statistics (Jenni & Loewenstein, 1997). This inconsistent valuation results in inefficient resource allocation and appears to be supported by substantial empirical research (Bergh & Reinstein, 2021; Caviola et al., 2020; Erlandsson et al., 2014; Friedrich & McGuire, 2010; Lee & Feeley, 2016; Loewenstein et al., 2006; Slovic, 2007; Small & Loewenstein, 2003).

The identifiable victim effect plays an important role in research and policy, with several highly cited articles and popular science books on the topic (e.g., Banerjee & Duflo, 2011; Bekkers & Wiepking, 2011; Singer, 2019; Slovic, 2007; Small et al., 2007). Further, charities often display images of identified victims, hoping that this increases charitable donations (e.g., <https://www.savethechildren.org.uk/>).

Given the importance of this phenomenon, it is concerning that several recent studies have questioned its robustness. For example, Hart et al. (2018) failed to find

support for the identifiable victim effect in donation behaviour and policy support. Recently, Moche and Västfjäll (2021) also failed to replicate the effect across three well-powered studies. A field experiment also failed to provide evidence for the effect (Lesner & Rasmussen, 2014). These failed replications are surprising given that other high-powered studies did find evidence for the identifiable victim effect (Caviola et al., 2020; Galak et al., 2011; Sudhir et al., 2016).

All of the studies cited above employ slightly different designs and can thus be considered conceptual replications. Reliance on conceptual replications can be problematic because when conceptual replications fail, it can be argued that the differences in methodology are the explanation for the different results (Chambers, 2017, p. 16). This may interact with publication bias, resulting in successful conceptual replications being published at a higher rate than unsuccessful ones.

This chapter, therefore, assesses the credibility of evidence for the identifiable victim effect more directly using a combined approach of bias-corrected meta-analysis and replication. We first reanalyse a widely cited meta-analysis on the topic (Lee & Feeley, 2016) and then attempt a direct replication of one of the key studies (Small et al., 2007).

2.1 Reanalysis of a Meta-Analysis on the Identifiable Victim Effect Suggests the Need to Revisit the Phenomenon

Lee and Feeley (2016) conducted a meta-analysis that summarized 41 effects from 22 experiments ($n = 15,967$) on the identifiable victim effect and found a ‘significant yet modest IVE [identifiable victim effect]’ (Lee & Feeley, 2016, p.199), referring to an aggregated effect of $r = .05$. However, there is reason to believe that this effect might be even weaker when publication bias is accounted for: the three highest powered studies in the dataset show effects that are almost zero, including one study with 12802 participants ($r = 0.004$). Lee and Feeley examined the possibility of publication bias using visual inspection of funnel plots. However,

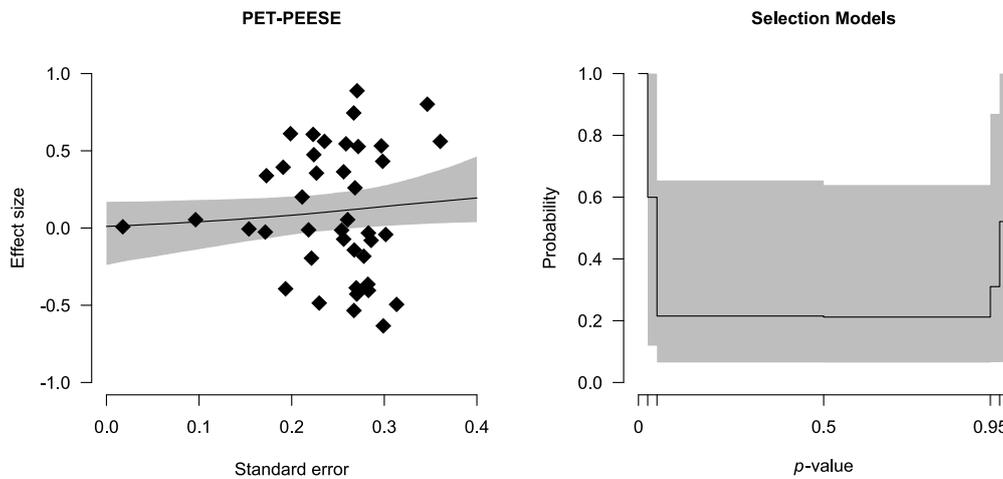
this approach does not perform well under some conditions like high heterogeneity (Bartoš, Maier, Quintana, & Wagenmakers, 2022; Bartoš, Maier, Wagenmakers, et al., 2022; Carter et al., 2019; Hong & Reed, 2021; Kvarven et al., 2020; Lau et al., 2006; Maier et al., 2022), which is present in Lee and Feeley’s meta-analysis ($QT[40] = 104.65, p < .001, I^2 = 61.8\%$). We therefore employ Robust Bayesian Meta-Analysis (RoBMA) to reanalyse the data by Feeley.

RoBMA quantifies evidence using Bayes factors. Bayes factors compare the likelihood of the data under competing models (in our case, the alternative hypothesis in comparison to the null hypothesis). In this chapter we report BF_{01} . In other words, Bayes factors that have the null in the numerator and the alternative in the denominator, and denote evidence in favour of the null hypothesis. As a rule of thumb for Bayes factors with the null in the numerator, Bayes factors between 1 and 3 are often regarded as weak evidence for the null, Bayes factors between 3 and 10 are often regarded as moderate evidence for the null, and Bayes factors larger than 10 are often regarded as strong evidence for the null (Jeffreys, 1939; M. D. Lee & Wagenmakers, 2013; Wasserman, 2000). However, we caution that these rules of thumb should merely aid interpretation and not be taken as absolute thresholds. Bayes factors are continuous measures of the strength of evidence, and any discretization inevitably results in loss of information.

When applying RoBMA to the data by Lee and Feeley (2016), we found moderate evidence for publication bias ($BF_{01} = 0.11$) and strong evidence for the absence of the average effect ($BF_{01} = 14.93$), with a model-averaged mean effect size estimate of $r = 0.002$ (95% CI [0, 0.004]). In addition, we find weak evidence against heterogeneity ($BF_{01} = 1.24$).

Due to the lack of publication bias correction methods that can accommodate a three-level structure, we accounted for the dependency between multiple estimates from the same study by using only the most precise estimate within each experiment. We further conducted a robustness check by selecting one estimate randomly within each study and bootstrapping 500 times. Using the median of these bootstraps, this analysis comes to the same conclusions regarding evidence for publi-

Figure 2.1: Footprint of Publication Bias in Lee and Feeley (2016)



Note. The left panel shows the PET-PEESE regression line (i.e., the relationship between effect sizes and standard errors) and the right panel shows the relative publication probabilities based on the selection models. The left panel displays a regression line of effect sizes on standard errors, the intercept of this line indicates the hypothetical estimate of a study with infinite precision; we can see that it is very close to 0. The right panel displays estimates for the relative publication probabilities of nonsignificant studies in comparison to significant studies model averaged across the different selection models included in RoBMA.

cation bias and evidence for an effect. Unlike the main analysis we find moderate rather than weak evidence against heterogeneity. In addition, as funnel plot based methods are sometimes criticized for finding bias for reasons other than publication bias (Lau et al., 2006; Maier et al., 2022), we also reanalysed the meta-analysis using only the selection models in RoBMA. This led to the same conclusions. We plotted the pattern of bias in Lee and Feeley in Figure 2.1. The left panel shows the regression line of effect sizes on standard errors. This relationship indicates that studies with smaller standard errors show smaller effects, a pattern that is indicative of publication bias. The right panel shows the relative publication probabilities for nonsignificant in comparison to significant p -values. This panel indicates that nonsignificant studies ($p > .05$) are considerably less likely to be published than significant studies.

2.2 Replication of Small et al. (2007)

2.2.1 Background

Given the substantial evidence for publication bias documented in the meta-analytic work, and evidence against the effect when this bias is accounted for, we proceed to replicate one of the seminal papers on this effect. We chose Studies 1 and 3 of Small et al. (2007) for replication due to the article's considerable impact. At the time of writing (May 2025), there were 1425 Google Scholar citations of the target article. Beyond the direct citation count, Small et al. (2007) have influenced several other highly cited articles (> 1000 times at the time of writing; e.g., Slovic, 2007; Bekkers and Wiepking, 2011) and popular science and philosophy books such as 'The Life You Can Save' (Singer, 2019) and 'Poor Economics' (Banerjee & Duflo, 2011), which have guided both research and policy.

To our knowledge, there has been one direct replication of Small et al. (2007): a Spanish language unpublished doctoral thesis failed to find support for the results of Study 1 (Charris, 2018). However, Charris (2018) only found weak evidence against the effect in a Bayesian analysis and no evidence for the null using the TOST procedure to test for equivalence (e.g., Lakens et al., 2018). Charris (2018) concluded that his study lacked statistical power and does not allow rejecting the identifiable victim effect.

2.2.2 Small et al. (2007): Hypotheses and Findings

Small et al. (2007) proposed that thinking analytically about the value of lives reduced giving to an identifiable victim but not to statistical victims. They also suggested that implicitly inducing analytical reasoning about the value of lives reduced donations to an identifiable victim but not to statistical victims. They conducted four experiments, and the current replication focused on Studies 1 and 3.

In Study 1, participants were randomly assigned to one of two conditions, with the intervention group learning about the identifiable victim effect from previous research (explicit learning condition), whereas another served as a control group. They were further randomly assigned to either the statistical victim condi-

tion, in which they read ‘factual information taken from the Save the Children website (<http://www.savethechildren.org/>) about the problems of starvation in Africa’ (Small et al., 2007, p.146), or to the identifiable victim condition, in which they received a brief description of an African girl from the Save the Children website. They were then instructed to donate any five one-dollar bills received earlier from a survey. After their donation, participants rated different affective reactions they experienced towards the described victim(s). These items included feeling upset, touched, sympathetic, and morally responsible, as well as the perceived appropriateness of donating to help the described victims.

To summarize, their Study 1 design was a 2 (Identifiability: identifiable vs. statistical) x 2 (Explicit Learning: intervention vs. control) between-subjects factorial design. Their results showed that in the control condition without the intervention, donations to the identifiable victim were higher than donations to statistical victims. However, the pattern was different for the participants who were assigned to the explicit learning intervention conditions and learned about the identifiable victim effect before asking to donate, with the donations being similar towards the identifiable victim compared to statistical victims. Surprisingly, this occurred because donations to the identifiable victim had been reduced rather than because donations to the statistical victim had increased.

In Study 3, Small et al. (2007) further studied the effect of implicit learning by adding a third identifiability condition, a joint condition (also referred to as ‘implicit learning condition’) that included both a picture of the single victim and general victim statistics, resulting in a three conditions design (identifiable vs. statistical vs. joint). The donation in this joint condition was intended for the described identified victim. The presentation of victim statistics was meant to implicitly eliminate the identifiable victim effect in the joint condition arguably because providing statistics alongside the victim reminds the potential donor of the many people who would not receive help. Small et al. (2007) found support for implicit learning, as donations to the identified victim were lower in the joint condition compared to the identifiable victim condition.

2.2.3 Method

Overview of Replication Study

In this replication, we merged Studies 1 and 3 in Small et al. (2007) into a single experimental design to study both the explicit and implicit ways of debiasing the identifiable victim effect. Our study was a 3 x 2 experimental design varying identifiability of the victim (identifiable victim, statistical victims, and joint - identifiable victim alongside statistical victims) and explicit learning (present or not). We summarise the design in Table 2.1. We preregistered the study at: <https://osf.io/dc9kb/>. Deviations from our preregistration are stated in the ‘Deviations from preregistration’ section of the supplementary materials and also at the appropriate places in the methods section. We provide additional details regarding the procedure in the ‘Procedure’ subsection in the supplementary materials and the Qualtrics survey is provided with the preregistration in the OSF folder.

Table 2.1: Replication and Extension: Experimental Design

		Identifiability (IV1; between-subject)		
		Identifiable victim condition	Statistical victim condition	Joint condition (Implicit Learning)
Explicit Learning	Explicit learning intervention	Identifiable Explicit	Statistical Explicit	Joint Explicit
(IV2; between- subject)	No intervention (Control)	Identifiable Control	Statistical Control	Joint Control

Note. Joint condition displayed both an identifiable victim and general victim statistics.

Participants

The study received ethics approval from the University of Hong Kong (EA1908020). A total of 1004 Amazon Mechanical Turk (MTurk) participants were recruited from a US sample using CloudResearch (Mean age = 39.4, SD = 12.4; 465 females, 533 males, 6 prefer not to say).

We collected as many participants as we could afford with the available funding. The full report of power analysis can be found in the supplementary materials under the section ‘Power analysis of the original study effect’ and indicates that for the lowest powered effect (the interaction between Explicit Learning and Identifiability), a sample size of 314 would be sufficient to achieve 95% power for the original effect size. In addition, sensitivity power analyses indicate that our sample size would have 95% power to detect a very small effect size of $\eta_p^2 = 0.012$ with an alpha level of .05.

Exclusion Criteria

We pre-registered to focus on the full sample for the main analysis and assess robustness to exclusions, with several exclusion criteria for the supplementary analyses: low English proficiency (scored lower than 4 on a scale of 0 to 6); not being serious in completing the survey (scored lower than 3 on a scale of 0 to 4); correctly guessed the hypotheses; already seen the survey before; failure to complete the survey or completed in less than a minute; and not from the United States.

Fifty-six responses met the exclusion criteria. We found no major differences between the pre- and post-exclusion results. As preregistered, we focused on the full sample for data analysis. We summarized the results after exclusion in the supplementary materials (‘Exclusion based on preregistration criteria’), with a comparison of the findings (‘Pre-exclusions versus post-exclusions’). The results of this analysis are very similar to the results presented here.

Manipulations

Explicit Learning. Participants were randomly assigned to either the explicit learning intervention condition or to the control condition. Participants in the explicit learning intervention condition were instructed to read a passage about prior research findings on the identifiable victim effect used in the original studies. In other words, they were taught about the phenomenon before the donation.

Identifiability. Participants were then randomly assigned to one of the three Identifiability conditions. Those in the identifiable victim condition read about a child from Zambia suffering from starvation, accompanied by a black-and-white photograph and a short description. Those in the statistical victim condition read about numerical victim statistics to illustrate the millions of people living in a similar plight to the child described in the identifiable victim condition. The joint condition was a combination of the previous two conditions, with the same Zambian child presented in a photo with a brief description, along with the victim statistics provided in the statistical victim condition.

Forced manipulation comprehension checks. To ensure reading and comprehension of the scenarios, we added checks that the participants had to answer correctly in order to be able to proceed to the next page that presented the dependent measures. This is a noted deviation from the target article's design, which we added to address concerns that online sample participants may not have read or were inattentive to the scenario and the manipulation.

Measures

Participants were then presented with the following continuation of the scenario: "Imagine that you have just earned \$5 US dollars and you are given an opportunity to donate any amount of the money to the organization Save the Children". They then indicate their hypothetical donations from 0 to 5 US\$ in increments of \$1 (\$0, \$1, \$2, \$3, \$4, or \$5). The donation was to the specific victim in the identifiable and joint conditions and to the anonymous group in the statistical victim condition. We use hypothetical donations rather than real donations for a number of reasons. First, this project was related to a different replication project we conducted in Majumder et al. (2023) in which we failed to replicate the identifiable victim effect demonstrated by Kogut and Ritov (2005) who showed the effect using hypothetical donations, as many other studies examining the identifiable victim effect have. We aimed to make the two replications as similar as possible in their dependent variables to allow one replication to possibly inform the other. Second, we thought

it best to ensure that the effect holds with simpler hypothetical donations before embarking on a more complex and costly real donation study. Mean donations are typically higher for hypothetical donations than for real donations (Bekkers, 2006); however, we are not aware of any evidence of mechanisms that result in differences between conditions when switching from real to hypothetical donations.

2.2.4 Results

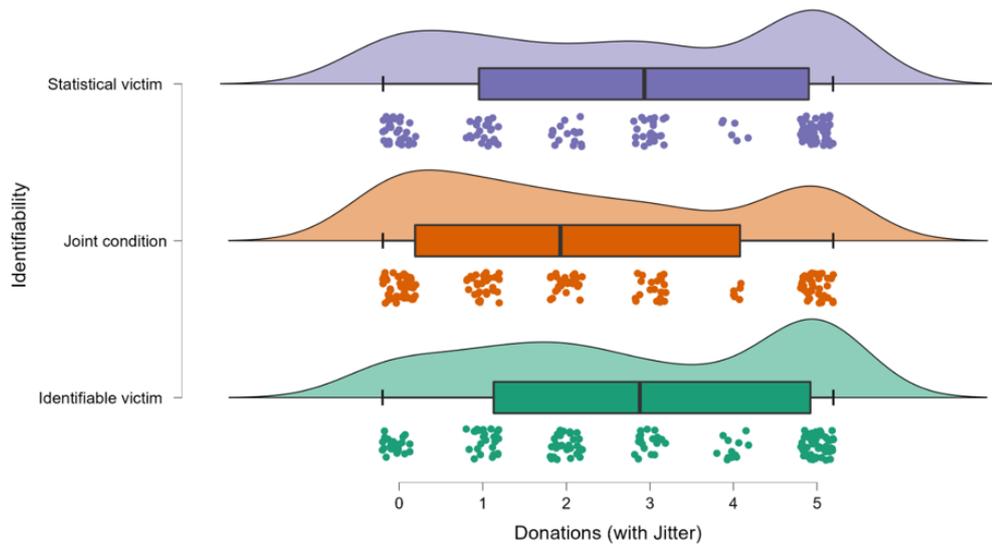
We followed and extended the analyses conducted by the target article. We provide a comparison of the statistical tests reported in the original study and the replication in the online supplementary materials.

Following the analyses conducted in Study 1 of Small et al. (2007), we carried out a 2 (Explicit Learning) \times 2 (Identifiability) two-way ANOVA (i.e., cells Identifiable-Explicit, Statistical-Explicit, Identifiable-Control, and Statistical-Control) to examine whether 1) people donate more when presented with an identifiable victim than when presented with a statistical victim; 2) the identifiable victim effect is weaker for people who were explicitly informed about the identifiable victim effect; 3) people that were explicitly informed about the identifiable victim effect tend to donate less than those uninformed about the effect. We plot donations by conditions (including joint condition) in Figure 2.2. We summarise the inferential tests of our replication in comparison to Small et al. (2007) in Table 2.2.

We supplemented the frequentist analyses with a Bayesian analysis to quantify evidence for the null. As prior distribution for the expected effect size, we use a Cauchy (0, 0.707), which is a common choice in Bayesian analysis. Because the Cauchy distribution is very fat-tailed, this prior gives a lot of mass to a wide range of plausible effect sizes while at the same time not reducing the ability to obtain evidence for smaller effects by much (Wagenmakers et al., 2020). As shown in Table 2.2, we found no support for the main effect of Identifiability, Explicit Learning, or their interaction, and absence of evidence in the Bayesian analysis, with similar donation amounts in the identifiable victim and statistical victim conditions. We therefore concluded failure to replicate the identifiable victim effect, and failure to replicate that explicitly learning about the effect impacted the effect itself.

Figure 2.2: Hypothetical Donations: Interaction of Identifiability and Explicit Learning

A: Explicit Learning Intervention



B: Control (No Intervention) Condition

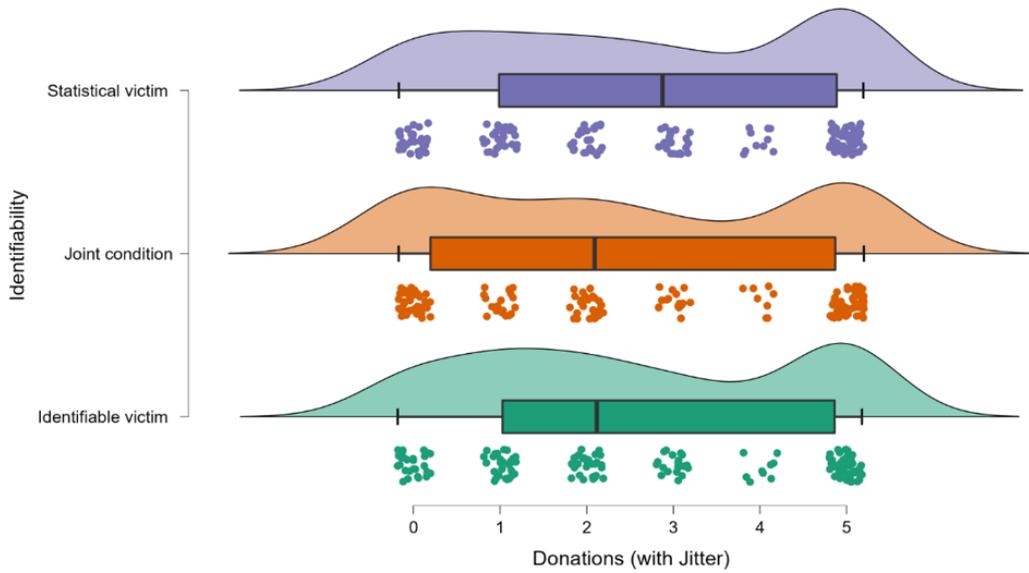


Table 2.2: Hypothetical Donations: Statistical Tests for Identifiability and Explicit Learning

	F	p	BF_{01}	η_p^2	95% CI
Identifiability					
<u>Without joint condition [S1]</u>					
Identifiable (Explicit & Control) vs. Statistical (Explicit & Control)					
Target article	6.75	< .05	N/A	.06	[.00, .15]
Replication	0.01	.923	11.57	.00	[.00, .003]
<u>With joint condition [E]</u>					
Identifiable (Explicit & Control) vs. Statistical (Explicit & Control) vs. Joint (Explicit & Control)					
Replication	3.91	.020	1.77	.01	[.00, .021]
Interaction: Identifiability and Explicit Learning					
<u>Without joint condition [S1]</u>					
Identifiable-Explicit vs. Statistical-Explicit vs. Identifiable-Control vs. Statistical-Control					
Target article	5.32	< .05	N/A	.04	[.00, .14]
Replication	0.654	.419	6.30	.001	[.000, .011]
<u>With joint condition [E]</u>					
Identifiable-Explicit vs. Statistical-Explicit vs. Joint-Explicit vs. Identifiable-Control vs. Statistical-Control vs. Joint-Control					
Replication	1.48	.228	12.74	.003	[.000, .012]
Explicit learning intervention					
<u>Without joint condition [S1]</u>					
Explicit (Identifiable & Statistical) vs. Control (Identifiable & Statistical)					
Target article	4.15	< .05	N/A	.03	[.00, .12]
Replication	0.89	.346	7.51	.00	[.000, .012]
<u>With joint condition [E]</u>					
Explicit (Identifiable-, Statistical, & Joint) vs. Control (Identifiable, Statistical, & Joint)					
Replication	0.005	.943	14.15	.00	[.000, .002]
Implicit learning and Identifiability					
<u>Without explicit learning [S3] *</u>					
Identifiable (Control) > Statistical (Control) \neq Joint (Control)					
Target article	5.67	< .01	N/A	.07	[.01, .15]
Replication	0.61	.541	25.03	.00	[.00, .015]

Note. ANOVA tests. $N = 1004$. CI = confidence interval. N/A = could not be recalculated. BF_{01} denotes the Bayes factor in favor of the null. Bayes factors based on Cauchy prior with $r_{scale} = 0.707$. η_p^2 for original study recalculated based on F-statistics and degrees of freedom. [S1] mirrors the target article's Study 1. [S3] mirrors the target article's Study 3. [E] indicated an extension. * indicates analysis was not pre-registered.

We ran an additional more complex version of the analysis above, which included a comparison to the joint condition (which was added in the target article's Study 3). This was only possible because of our unified design combining replications of the target article's Studies 1 and 3. We conducted a 2 (Explicit Learning) \times 3 (Identifiability) two-way ANOVA to examine if the provision of additional quantitative information together with an identified victim would debias the identifiable victim effect.

We found no support for the main effect of Explicit Learning and no interaction effect of Identifiability and Explicit Learning. We found some support for main effect of Identifiability, $F(2, 998) = 3.91, p = .02, \eta_p^2 = .008$. 95% CI [.000, .021], though Bayesian analysis indicates weak support for the null ($BF_{01} = 1.77$). The different conclusions from the two analyses can be explained by the large sample

that increases the likelihood of significant p-values, even when the evidence is low from a Bayesian perspective (Maier & Lakens, 2022).

To better understand the Identifiability main effect, we also examined the post-hoc comparisons comparing the different Identifiability conditions with Bonferroni correction. We found no support for differences between statistical and identifiable victim conditions, $t(998) = 0.097, p = 1.00, d = 0.01, 95\% \text{ CI } [-0.16, 0.14]$, and near threshold for the comparison between identifiable and joint, $t(998) = 2.37, p = .053, d = 0.18, 95\% \text{ CI } [0.03, 0.34]$ with donations slightly lower in the joint condition. We found support for differences between the statistical and the joint condition $t(998) = 2.46, p = .041, d = 0.19, 95\% \text{ CI } [0.04, 0.34]$. Given the weak near-threshold unexpected effect, we caution against over-interpretation of the Identifiability main effect or the contrasts.

Finally, we conducted an analysis mirroring the analyses of Study 3 in Small et al (2007) by not including the explicit learning intervention conditions. Although this was conducted by the target article, it was not included in the pre-registration, which was focused on the unified design and included the explicit conditions (see below). We therefore labeled this analysis exploratory. We found no support for an implicit learning effect, $F(2, 499) = 0.61, p = .541, \eta_p^2 = .002, 95\% \text{ CI } [.00, .02]$, and strong evidence against the effect in a complementary Bayesian analysis ($\text{BF}_{01} = 25.03$). Therefore, we did not conduct any follow-up tests comparing differences between specific cells.

2.3 Discussion

This chapter revisited the Identifiable Victim Effect using both publication bias-adjusted meta-analysis and a replication of a seminal paper. We found evidence against the IVE in a reanalysis of a large meta-analysis (Lee and Feeley, 2016) using novel publication bias adjustment methods. In our replication and extension of Small et al. (2007), we found no support for the identifiable victim effect and Bayesian analyses indicated evidence against an effect.

We caution that our results should not be considered a ‘final word’ on this effect

but rather a motivation for future replication efforts in the form of high-powered registered reports examining hypothetical donations, donation intent, real money donations, and associated perceptions such as perceived impact. In addition, we see many promising theoretical directions for further work in this area and possibilities for rethinking and reframing the original theory.

2.3.1 Identifiable Victim Effect or Scope Insensitivity?

Majumder et al. (2023) recently reported a failed replication of Kogut and Ritov (2005) and suggested that the identifiable victim effect may be reframed: instead of larger donations towards an identifiable victim, the effect might be viewed as similar donations towards an identifiable victim as a group of unidentified or statistical victims with no donation adjustment per group size. This cognitive phenomenon is usually discussed under the term ‘scope insensitivity’, and describes that people do not value a good (here helping children in need) in proportion to its scope or size (Baron & Greene, 1996; Desvousges et al., 1993; Kahneman & Knetsch, 1992). Scope insensitivity has also been shown to be a factor in charitable giving (Hsee et al., 2013; Västfjäll and Slovic, 2020, see also Chapter 4 of this thesis) and has been discussed as a reason for neglecting to help save human lives, for example, in the context of genocides (Cameron & Payne, 2011; Dickert et al., 2012, 2015; Slovic & Västfjäll, 2010b). Given the results from this study it is possible that the Identifiable Victim Effect should be reconsidered in terms of scope insensitivity, a topic that is addressed in Chapter 4 of this thesis.

2.3.2 Evidence for Irrational Decision Making?

It is unclear whether this form of equal contribution to identifiable versus statistical victims can be considered evidence for irrational decision making in the context of identifiable victims. On the one hand, as Majumder et al. (2023) argued, the larger group of victims should elicit more empathic concern, distress, and consequently, willingness to contribute. Not observing this pattern violates the principle of proportionality (i.e., larger issues should be tackled with more resources). On the other hand, from a cost-effectiveness perspective, it makes sense to contribute more

where the donation is most effective rather than where the problem is biggest. In our study, we found that participants did not necessarily perceive a hypothetical donation to the larger group as more impactful, but rather that they may in fact consider donations to the identifiable victim more impactful (see Maier, Wong, and Feldman, 2023, for the detailed write-up including this finding). People might also perceive donating to the identified victim as more impactful due to proportion dominance. In other words, they may donate less to statistical victims, given that they perceive a lower impact of their contribution when they can only help a smaller proportion of affected individuals (e.g., Erlandsson et al., 2014).

Therefore, effectiveness-based reasoning would imply the opposite compared to the principle of proportionality – donating more to the identified victim. A potential explanation of the null effects in our study would be that participants apply both reasoning based on proportionality and based on effectiveness, and the two cancel each other out, resulting in an overall null effect. Future research may measure participants' effectiveness focus and tendency to allocate resources based on proportionality to directly investigate how these two factors affect donations to the identifiable victim.

2.3.3 Limitations and More Future Directions

A core limitation that may explain the discrepancy between the results of the original studies and our replication is our adjustment from real to hypothetical donations. In Small et al. (2007), participants received money as a reward after filling in an unrelated survey about the use of various technology products. Participants then received a blank envelope and a charity request letter to decide how much they would be willing to donate. Answering the unrelated technology survey allowed participants to assess how much effort they invested to earn money, making it easier to grasp the subjective value of the money than in our study. Second, given this cover story, participants may not have realized that the experimenters were investigating their donation behaviour. Third, participants might donate differently with real in comparison to imaginary money, as they would, for instance, likely deliberate more when making choices involving real donations.

In our replication, we asked the participants to imagine they had just earned \$5 and how much of this they would like to give to the corresponding victims. Generosity reflected in the hypothetical donation is usually higher than that expected in the original studies (Bekkers, 2006). Though a direct comparison between the two studies is problematic given the passing of time and the very different measures, looking at the raw numbers in our replication, people indicated higher hypothetical donations, compared to the real donations reported in the target article. However, we note that our conclusions do not depend on average donations but on the differences between conditions. We are not aware of any evidence that would suggest that these effects stand a better chance of working with real donations than they do in hypothetical scenarios. Nevertheless, a replication in a field setting or an experiment with real donations would be valuable in the future, though we recommend adjusting expectations and taking into account that observed effects might be much weaker than initially thought.

Second, we made additional adjustments and also added forced comprehension checks, to ensure that participants read and understood the hypothetical donation situation and choice. It is possible that this may have somehow impacted participants' responses since they might disrupt feelings of empathy. In addition, participants may believe that the information about the identified victim effect was supplied to them in order to answer the comprehension checks rather than in order to use it in the subsequent donation task. However, we note that if the effect was indeed affected by such factors, it may indicate that the initial demonstrations were at least partially motivated by socially desirability responding (McKenzie et al., 2018), and/or that the effect is more contextual, weaker, and less robust than initially thought.

Third, our study was conducted online rather than in person (as in Small et al., 2007). On the one hand, this difference may be considered a strength, as the online data collection allowed us to collect a larger and broader sample than would have been possible in a lab study. On the other hand, the increased anonymity in online settings could reduce participants' willingness to donate, even though it is

less clear how this would affect the differences between conditions. This research was also conducted during the Covid-19 pandemic, which might have affected participants' financial status and their psychological state more broadly. These two factors might have resulted in our participants having little money for donations or being pre-occupied with financial and existential concerns. Hypothetical donations, therefore, may have been limited by participants resource constraints, or their 'mental account' of how much they are willing to contribute to donation tasks (Sussman et al., 2015; Thaler, 1985, 1999).

2.3.4 Conclusion

This chapter finds no support for the identifiable victim effect both in a bias-adjusted meta-analysis as well as in a replication of Small et al. (2007). We emphasize that this chapter should not be considered conclusive evidence against the identifiable victim effect, given the differences in the experimental setup. Instead, we believe that the failure to find the effect on hypothetical donations in combination with the publication bias-adjusted meta-analysis constitutes a cautionary note. Our work thus shows a pressing need for more replications with real donations in the form of registered replication reports (Chambers, 2013), ideally conducted as adversarial collaborations between proponents and critics of the identifiable victim effect.

Reference to Journal Article: Maier, M.,* Wong, Y.*, & Feldman, G.* (2023). Revisiting and rethinking the identifiable victim effect: Replication and extension of Small, Loewenstein, and Slovic (2007). *Collabra: Psychology*, 9(1).<https://doi.org/10.1525/collabra.90203>. (The chapter in this thesis is considerably shorter than the article, as the focus of Part 1 is mostly on publication bias adjustment, and the full paper, including more than 10 tables, would be beyond the scope of this thesis.)

Data and Code Availability: Data and analysis code are available at <https://osf.io/n4jkh/>.

Supplemental Information: Supplementary Materials are available at <https://tinyurl.com/f5zbkyrf>.

Author contributions: Maximilian Maier built on the thesis work by Yik Chun,

verified all analyses, added additional analyses (Bayesian), the RoBMA reanalysis of Lee and Feeley (2016), new visualizations, and wrote an initial journal submission manuscript. Yik Chun Wong conducted the replication as part of her thesis. Gilad was the advisor for the thesis. Gilad supervised each step in the project, conducted the preregistrations, and ran data collection. Maximilian and Gilad finalized the journal submissions, revised and responded to peer review.

Chapter 3

Nudging

In the previous chapter, I found no evidence that providing people with information about an identifiable victim increases willingness to donate. This chapter will broaden the scope and assess the evidence for nudges (i.e., behaviour change interventions that do not involve incentives) more broadly. Nudging is one of the most widespread applications of behavioural science to public policy. Nudge theory postulates that small changes in choice architecture substantially influence real-world decision making (Thaler & Sunstein, 2009). Unlike most other forms of influence, nudges maintain freedom of choice by not restricting choice options. This is a key benefit over more coercive policy measures, which has led to widespread interest in nudges both in academia and government research units ('Nudge Units'). However, despite its considerable promise, nudging has been criticised for its limited evidence base (e.g., Lin et al., 2017). This chapter, therefore, revisits the evidence base for nudging using publication bias correction methods. In the first part of this chapter, I reanalyse a large meta-analysis on nudging (Mertens et al., 2022) to investigate publication bias on nudging in the academic literature, while in the second part, I take a closer look at work by 'Nudge Units', such as the Behavioural Insights Team.

3.1 No Evidence for Nudging After Adjusting for Publication Bias

Mertens et al. (2022) provide a comprehensive assessment of the evidence for nudging in a timely meta-analysis with 200 studies reporting over 440 effects

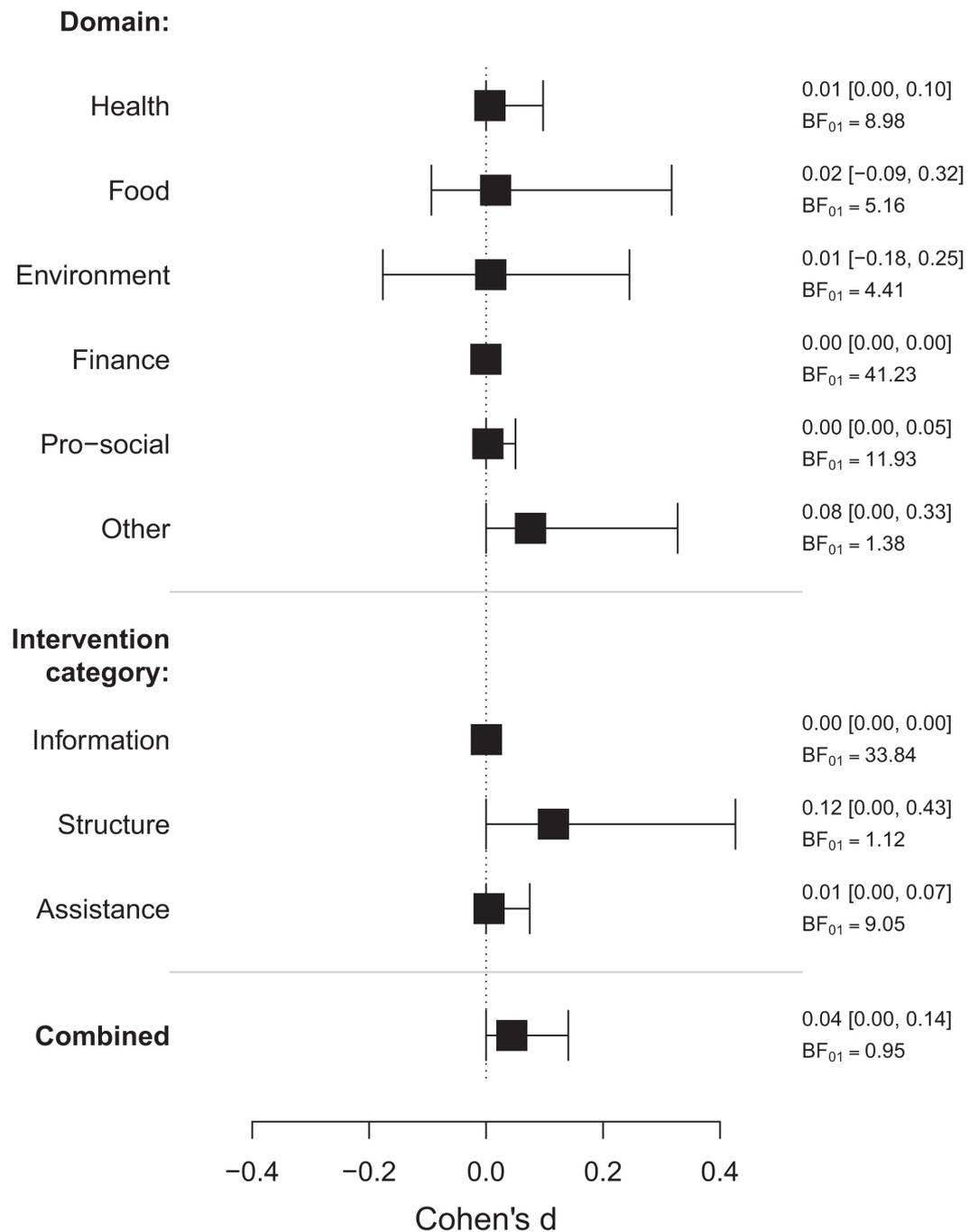
($n = 2,148,439$). They include a wide range of choice architecture interventions spanning different types of interventions (focusing on decision structure, decision assistance, and decision information) and domains (health, food, environment, finance, pro-social, and other; for more details on the methodology see Mertens et al., 2022, pp. 8–9). Mertens et al.’s headline finding is that “choice architecture [nudging] is an effective and widely applicable behavior change tool” (p. 8). We propose their finding of “moderate publication bias” (p. 1) is the real headline; when this publication bias is appropriately corrected for, no evidence for the effectiveness of nudges remains (Figure 3.1).

Mertens et al. find significant publication bias, through Egger regression. Their sensitivity analysis (Vevea & Woods, 2005) indicates that the true effect size could be as low as $d = 0.08$ (if publication bias is severe). Mertens et al. argue that severe publication bias is only partially supported by the funnel plot and proceed largely without taking publication bias into account in their subsequent analyses. However, the reported Egger coefficient ($b = 2.10$) is “severe” (Doucouliagos & Stanley, 2013).

In contrast, Robust Bayesian Meta-Analysis (RoBMA) (Maier, Bartoš, & Wagenmakers, 2023), avoids an all-or-none debate over whether or not publication bias is “severe” and has several advantages over funnel plot-based methods (outlined in Chapter 1). We therefore apply RoBMA to the data of Mertens et al. (2022).

Table 1 compares the unadjusted results to the publication bias–adjusted results. Since publication bias–corrected three-level selection models are computationally intractable, we analysed the data in two ways: 1) ignoring the three-level structure (column 2) and 2) using only the most precise estimate from studies with multiple results (column 3). Strikingly, there is an absence of evidence for an overall effect and evidence against an effect in the “information” and “assistance” intervention categories, whereas the evidence is undecided for “structure” interventions. When using only the most precise estimates, we further find evidence against an effect in most of the domains, apart from “other,” “food,” and “prosocial” (the

Figure 3.1: Correcting for Publication Bias Suggests No Evidence for the Mean Effect of Nudging



Note. RoBMA_{PSMA} model-averaged posterior mean effect size estimates with 95% credible intervals and Bayes factors for the absence of the effect for the combined sample or split by either the domain or intervention category (ignoring the clustering of SEs). BF₀₁ quantifies evidence for the null hypothesis. BF₀₁ larger than one corresponds to evidence in favor of the null hypothesis, and BF₀₁ lower than one corresponds to evidence in favor of the alternative hypothesis (evidence for the alternative hypothesis can be obtained by reciprocating the Bayes factor; BF₀₁ = 1/BF₀₁). As a rule of thumb, Bayes factors between 3 and 10 indicate moderate evidence, and Bayes factors larger than 10 indicate strong evidence.

evidence is indecisive) and weak evidence for the overall effect.¹ However, all intervention categories and domains apart from “finance” show evidence for heterogeneity, which implies that some nudges might be effective, even when there is evidence against the mean effect. Finally, we find strong evidence for publication bias across all subdomains ($BF_{pb} > 10$), apart from food, when using only the most precise estimates ($BF_{pb} = 2.49$).

Table 3.1: Comparison of Unadjusted and Adjusted Effect Size Estimates for All Studies and for Subsets of Studies Based on Different Categories or Domains.

	Random effects	RoBMA_{PSMA}	RoBMA_{PSMA} (precise)
Combined	0.43 [0.38, 0.48] $t(333) = 16.51$	0.04 [0.00, 0.14] $BF_{01} = 0.95$	0.11 [0.00, 0.24] $BF_{01} = 0.31$
Intervention category			
Information	0.25 [0.19, 0.30] $t(88) = 8.79$	0.00 [0.00, 0.00] $BF_{01} = 33.84$	0.00 [0.00, 0.07] $BF_{01} = 10.57$
Structure	0.58 [0.50, 0.66] $t(186) = 13.93$	0.12 [0.00, 0.43] $BF_{01} = 1.12$	0.23 [0.00, 0.49] $BF_{01} = 0.33$
Assistance	0.22 [0.15, 0.29] $t(65) = 6.42$	0.01 [0.00, 0.07] $BF_{01} = 9.05$	0.01 [0.00, 0.12] $BF_{01} = 8.00$
Domain			
Health	0.31 [0.22, 0.39] $t(64) = 7.03$	0.01 [0.00, 0.10] $BF_{01} = 8.98$	0.02 [0.00, 0.19] $BF_{01} = 3.53$
Food	0.66 [0.52, 0.81] $t(81) = 9.01$	0.02 [-0.09, 0.32] $BF_{01} = 5.16$	0.27 [0.00, 0.64] $BF_{01} = 0.55$
Environment	0.48 [0.37, 0.58] $t(56) = 9.16$	0.01 [-0.18, 0.25] $BF_{01} = 4.41$	0.00 [-0.44, 0.34] $BF_{01} = 3.05$
Finance	0.23 [0.15, 0.31] $t(34) = 6.08$	0.00 [0.00, 0.00] $BF_{01} = 41.23$	0.00 [0.00, 0.00] $BF_{01} = 30.95$
Prosocial	0.32 [0.22, 0.42] $t(38) = 6.36$	0.00 [0.00, 0.05] $BF_{01} = 11.93$	0.05 [0.00, 0.27] $BF_{01} = 1.89$
Other	0.40 [0.29, 0.50] $t(55) = 7.66$	0.08 [0.00, 0.33] $BF_{01} = 1.38$	0.04 [-0.22, 0.40] $BF_{01} = 2.45$

Note. First column: Random effects meta-analysis estimates with 95% CI based on clustered SEs, all P values < 0.001. Second and third columns: RoBMAPSMA model-averaged posterior mean effect size estimates with 95% credible intervals and Bayes factor for the presence of the effect ignoring the clustering of SEs or using the most precise estimates (precise). Results differ slightly from the moderator analysis presented in the article because we analysed each subfield separately to allow 1) testing for the presence of the effect in each category/domain in the Bayesian framework, and 2) publication bias to operate differently in different subdomains.

We conclude that the “nudge” literature analysed in Mertens et al. (2022) is characterized by severe publication bias. Contrary to Mertens et al., our Bayesian analysis indicates that, after correcting for this bias, no evidence remains that

¹We also reanalysed the data by including only models of selection for statistical significance, confirming our results.

nudges are effective as tools for behaviour change.

Reference for Journal Article: Maier, M.*, Bartoš, F.*, Stanley, T.D., Shanks, D.R., Harris, A.J.L., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, 119(31), e2200300119. <https://doi.org/10.1073/pnas.2200300119>

Data and Code Availability: Data are available at <https://osf.io/ubt9a> and analysis code is available at <https://osf.io/svz6e/>.

Author contributions: M.M. and F.B. designed research; M.M. and F.B. performed research; M.M. and F.B. analysed data; and M.M., F.B., T.D.S., D.R.S., A.J.L.H., and E.-J.W. wrote the paper.

3.2 Exploring Open Science Practices in Behavioural Public Policy Research

Next to academic research, the popularity of nudges has motivated the creation of nudge units: government agencies or independent companies that evaluate different behavioural interventions to inform decisions on whether to roll them out more widely (more than 200 nudge units in more than 40 countries have been created to date; DellaVigna and Linos, 2022, Figure A1). Nudge units aim to deliver substantial policy benefits with comparatively small interventions (Halpern, 2015).

The UK Behavioral Insights Team (BIT), founded in 2010 and the oldest and largest behavioural insights team, has completed more than 1000 projects.² The BIT website lists 137 reports and 36 publications, usually produced in collaboration with government agencies. There are a number of success stories amongst these projects, where considerable real-world benefits appear to have been delivered. In one trial, for example, BIT used behavioural insights to design better tax reminder messages using social norms, leading to increased average payments (Hallsworth et al., 2017).³ BIT is a large multinational organization, with offices in multiple

²<https://web.archive.org/web/20240108153747/https://www.bi.team/about-us-2/who-we-are/>

³However, this effect failed to replicate in a different council (P. John, Blume, et al., 2018).

countries, including the UK, Canada, the United States, France, Australia, and Singapore. It was formed within the UK government but is now a social purpose organisation operating outside the government. In the US, the Office of Evaluation Sciences (OES) was established by a Presidential Executive Order in 2015 with the mission to rigorously test and incorporate behavioural insights into government agencies. OES has completed over 90 impact evaluations affecting the lives of millions of citizens.⁴ Compared with BIT, OES is a comparatively small team that operates within the US government. Crucially, behavioural science units use randomized controlled trials (RCTs)—the ‘gold standard of evaluation’. For example, BIT has completed more than 700 RCTs to date in many different countries.⁵ This adoption of RCTs has enhanced the evidence base for government policy.⁶

The nudge approach is not, however, without critics (Chater & Loewenstein, 2022). Two main objections are: 1) Despite the aforementioned success stories, overall evidence for the effectiveness of nudges in the academic literature is weak as shown in the previous section of this chapter (see also Bakdash and Marusich, 2022; Szaszi et al., 2022); 2) nudge-based interventions may detract from more systemic reforms (Chater & Loewenstein, 2022). These criticisms culminated in a recent manifesto for applying behavioural science (Hallsworth, 2023), proposing a variety of reforms and calling for ‘increased self-scrutiny’. Following these calls, we take a close look at the distribution of test statistics and safeguards against biased reporting in nudge unit trials. We argue that nudge units can enhance their current practices with specific improvements in the transparency of their trial registration, reporting, and data sharing.

⁴<https://web.archive.org/web/20240117111129/https://oes.gsa.gov/work/>

⁵<https://web.archive.org/web/20240108153747/https://www.bi.team/about-us-2/who-we-are/>

⁶For example the European Commission states about behavioural insights: ‘In practice, however, behavioural insights mainly contribute to the impact assessment process. This process consists in gathering and analysing evidence about the likely impacts of a planned policy.’ https://web.archive.org/web/20240110142327/https://knowledge4policy.ec.europa.eu/behavioural-insights/about-behavioural-insights_en

3.2.1 Exploring Potential Reporting Biases in Nudge Unit Trials Using Bias Correction Techniques

DellaVigna and Linos (DellaVigna & Linos, 2022) collected a large dataset of nudge unit interventions run by OES and BIT North America (126 randomized control trials covering 23 million individuals)⁷ and compared them to trials in academic journals to evaluate the shrinkage of effects when applied at scale. The comparison showed that the average impact of nudges reported in academic journals (8.7 percentage points increased take-up, a 33.4% increase over the average in the control condition) was larger than in trials run by OES and BIT (1.4 percentage points increased take-up, an 8.0% increase over the control condition). In line with the findings from the first section of this chapter, this was primarily attributed to selective publication and low statistical power in the academic studies. Although with smaller effect sizes, the nudge unit interventions were found to produce reliable, ‘sizable and highly statistically significant’ (p. 81) effects. Importantly, DellaVigna and Linos (DellaVigna & Linos, 2022) assumed no selective reporting in the nudge unit interventions because they obtained access to the comprehensive record of trials. In addition, they visually inspected the distribution of t -statistics and conducted a regression test for funnel plot asymmetry, testing both the relationship between minimum detectable effect and treatment effect, as well as between standard error and treatment effect. Both the visual inspection of t -statistics and the regression indicated no evidence for publication bias.⁸

However, while the comprehensive record of trials protects from publication bias (when a complete study is omitted), it does not necessarily protect from other forms of selective reporting (e.g., choosing which outcome variables to report or emphasize, or what covariates to include). Further, both visual inspection of funnel plots, as well as regression of effect sizes on standard errors, have been shown

⁷This is less than the number quoted in the introduction as many BIT trials have been conducted outside the US.

⁸In contrast, in online Appendix A2, DellaVigna and Linos (DellaVigna & Linos, 2022) show that published articles based on the nudge unit interventions suffer from the same pattern of publication bias as published academic papers.

in simulation studies and empirical examples to often have low power to detect reporting biases especially under high heterogeneity (Maier, Bartoš, & Wagenmakers, 2023; Terrin et al., 2005). Here, we therefore apply statistical techniques that are more suitable to test for potential reporting biases in the presence of heterogeneity (McShane et al., 2016) to the nudge unit dataset (i.e., 241 nudges from 126 trials, as collected by DellaVigna and Linos, 2022).

DellaVigna and Linos (2022) extend the standard meta-analytic framework by modelling the effect sizes as a two-component random effects meta-analytic mixture. This means that instead of assuming that all effects come from a single distribution, as is common in meta-analyses, their framework allows the effect sizes to come from two separate distributions. *Prima facie*, such an approach seems reasonable given the large differences between different behavioural interventions incorporated under the term ‘nudge’ (Szasz et al., 2022). For example, researchers might assume that effect sizes for nudges that change the default option are distributed differently from nudges with smaller effects.⁹ Here, we follow DellaVigna and Linos and take a data-driven approach to determine the appropriate number of distributions. Models assuming a single distribution (i.e., the standard meta-analytic random effects model) are compared to models with larger numbers of mixture components using model selection techniques to find the appropriate model. DellaVigna and Linos (2022) show that assuming all effect sizes come from a single distribution does not adequately describe the data. We come to the same conclusion when comparing single-component models and mixture models using BIC. For this reason, and to keep our analysis comparable to that of DellaVigna and Linos, we proceed with the mixture modelling approach.

To assess selective reporting via bias correction methods, we extended DellaVigna and Linos’ (DellaVigna & Linos, 2022) analysis in three ways. First, we allow for moderation by domain within the mixture model (i.e., different areas in which nudges may be used, such as work and education or healthcare, as classified in

⁹The difference between these types of interventions could also be modelled by including appropriate moderators. However, often researchers do not know all the relevant differences between nudge characteristics a priori. This is also the case in DellaVigna and Linos (2022), which found evidence for mixtures despite including different moderators in the analysis.

DellaVigna and Linos, 2022). This is important, as inclusion of appropriate study level covariates may explain some of the non-normal heterogeneity that would otherwise be captured by using multiple mixture components.

Second, we additionally specify mixture models that allow for selective reporting—selection models—and compare them to the normal models. Selection models, as we specify them here, include an assumption that null or backfire effects are suppressed within the distribution of effects reported. We include three types of selection models: (1) ones that assume that negative results are less likely to be published than positive results, (2) ones that assume that non-significant studies at $\alpha = .10$ are less likely to be published than significant studies, and (3) ones that assume that non-significant studies at $\alpha = .05$ are less likely to be published than significant studies. We used BIC-based Bayesian model averaging to combine the evidence across the three types of selection models (Hoeting et al., 1999; Raftery et al., 1995).

Third, we also allow further expansion to three-component mixtures. This may improve model fit further compared to the two-component results.¹⁰

Overall, we fit the following six models to the full dataset (more details on the specified models are provided in the supplementary materials):

1. a random effects meta-analytic model (normal model);
2. a two-component random effects meta-analytic mixture model (2-mixture), as in DellaVigna and Linos (2022);
3. a three-component random effects meta-analytic mixture model (3-mixture);
4. a random effects meta-analytic model with adjustment for selective reporting (selection model);

¹⁰Two reviewers suggested applying the RoBMA method (Bartoš, Maier, Wagenmakers, et al., 2022; Maier, Bartoš, & Wagenmakers, 2023). RoBMA like most other ‘out-of-the-box’ meta-analytic methods, assumes that effect sizes follow a single distribution. Extending RoBMA, and other meta-analytic methods, to mixture modelling is a non-trivial endeavour (computational tractability, convergence, parameterization, ...), therefore, we proceeded with analysis analogous to DellaVigna and Linos (DellaVigna & Linos, 2022).

5. a two-component random effects meta-analytic mixture model with adjustment for selective reporting (selection 2-mixture), following DellaVigna and Linos (2022);
6. a two-component random effects meta-analytic mixture model with adjustment for selective reporting (selection 3-mixture).

We estimate the models using the `optim()` optimization routine from the `optim` package in R (version 4.3.2; Windows 11; R Core Team, 2021).

Figure 3.2 visualizes the model-fit of the different models to the full dataset. When looking only at mixtures of one and two components (matched to DellaVigna and Linos), we find that the data are most in line with a model assuming a mixture of two normals and selective reporting (BIC weights: normal model, 0.000; normal 2-mixture, 0.042; selection model, 0.000; selection 2-mixture, 0.958). The figure also clearly indicates that a single normal distribution does not capture the data well, which suggests that different types of nudges are described by different distributions.¹¹ When we also allow extension to three parameter mixtures, we find somewhat weaker evidence for selective reporting; however, most weight is still given to the selection 3-mixture, a model that assumes selective reporting (BIC weights: normal model, 0.000; normal 2-mixture, 0.000; normal 3-mixture, 0.249; selection model, 0.000; selection 2-mixture, 0.000; selection 3-mixture, 0.750).

We can also make inferences about the type of publication bias based on which types of selection models received the highest weight. This shows that the lowest BIC was given to the three component selection model, which assumes that results with negative estimates (rather than nonsignificant results) are suppressed. This model has the lowest BIC when looking at one component models, and the second lowest when looking at two component models (with the $\alpha = .10$ model being slightly preferred). Overall, this suggests that selective reporting operates most strongly on suppressing backfire effects rather than on selection for $p < .05$.¹²

¹¹In line with DellaVigna and Linos, we use the mixture model here to obtain a better non-parametric approximation of the distribution shape. It would be an interesting project to develop further the theoretical interpretation of each of the components, but this lies beyond the scope of the present article.

¹²An important consideration for future research is how the mixture models and the publication

Regardless of our precise choice of model, however, overall we find that most weight is given to models that assume selective reporting. We next directly compared pre-analysis plans to publicly-available final reports. This enables us to identify instances where pre-analysis plans may allow for selective reporting and provide corresponding recommendations.

3.2.2 Pre-analysis Plans Leave Scope for Selective Reporting

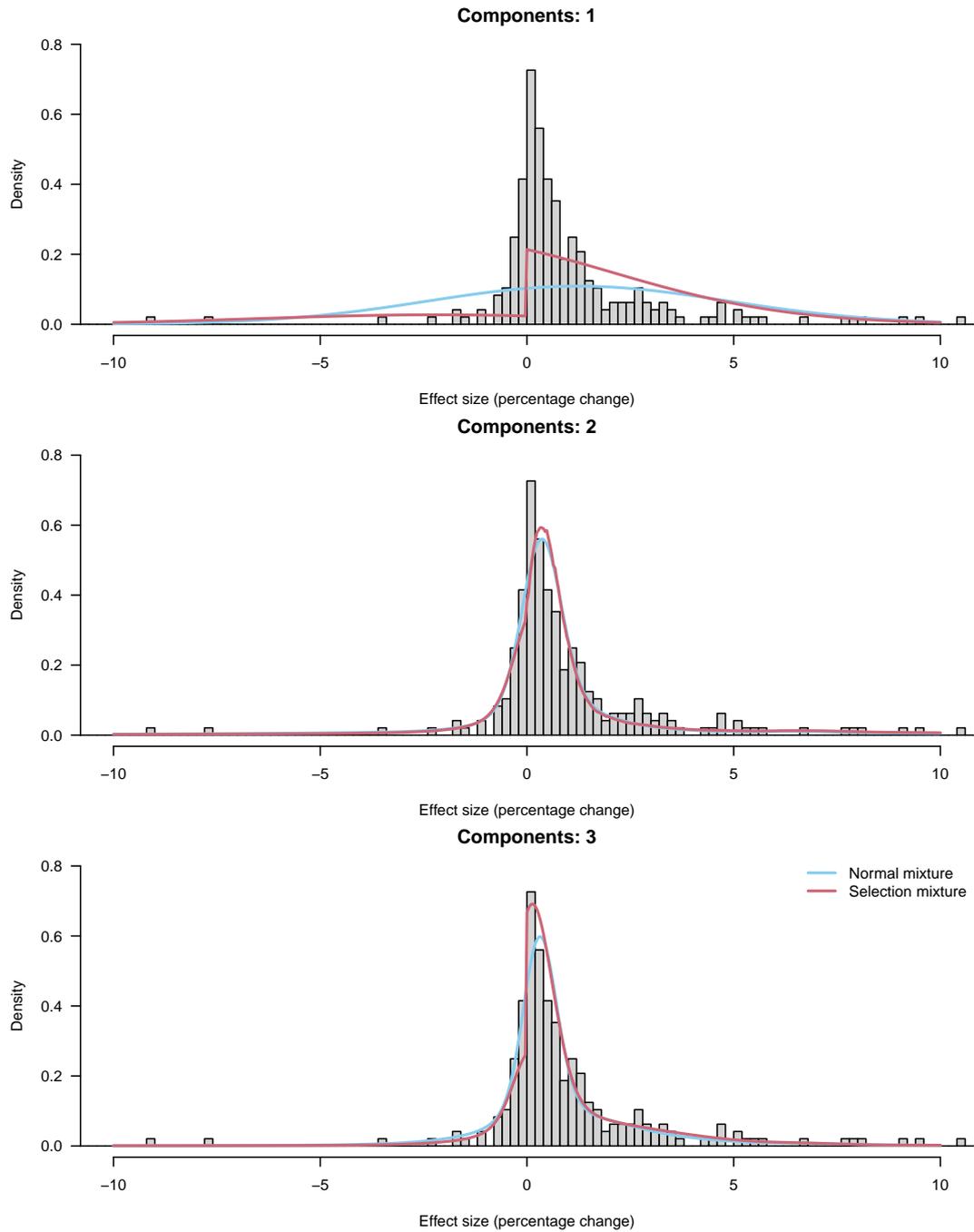
Both the BIT and OES document their intended analyses in pre-analysis plans. This is laudable, diminishes the scope for selective reporting, and enables evaluation of any deviations from such plans (if they are shared publicly). However, previous research comparing trial protocols or pre-registrations to corresponding published journal articles indicates that selective *reporting* is still possible, even without selective *publication* (e.g., in economics: Brodeur et al., 2022; in medicine: G. Li et al., 2018; in psychology: Claesen et al., 2021). Selective reporting practices include choosing which outcome variables to report or emphasize and what covariates to include. These practices can be unintentional - it is easy for any researcher to convince themselves that the analysis with covariate A is ‘most appropriate’ once knowing the outcome, without recognising the potential for bias in such a decision (Simmons et al., 2011).

Below, we investigate whether the pre-analysis plans of trials run by nudge units allow for selective reporting. We (a) evaluate how detailed the pre-analysis plans are, and whether they cover all relevant researcher degrees of freedom; and (b) compare pre-analysis plans to published reports to assess selective reporting. While we were unable to obtain the pre-analysis plans from BIT (in the UK or US), despite taking a variety of steps,¹³ OES trial protocols are publicly available. We searched

bias models may interact. Including the mixture model does weaken the evidence for selective reporting (as visible in the BIC differences in Appendix 3) and it may be that the mixtures can approximate patterns of selective reporting to some extent.

¹³First, we contacted the head of US BIT to ask for the protocols. Second, we tried to obtain the protocols via DellaVigna and Linos, who recommended contacting BIT directly. Third, we contacted the UK BIT through a form on their website and received an initial response, but this did not follow through to sharing the pre-analysis plans. Fourth, we tried to obtain the protocols through a Freedom of Information request at https://web.archive.org/web/20240117111819/https://www.whatdotheyknow.com/request/trial_protocols_behavioural_insi/#incoming-2143267 and https://www.whatdotheyknow.com/request/trial_protocols_behavioural_insi/#incoming-2143267 and after

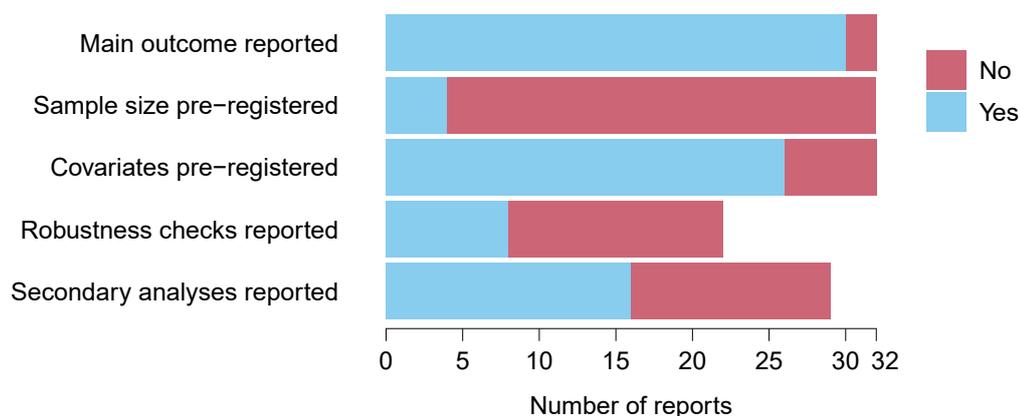
Figure 3.2: Distribution of All Effect Sizes and Visualization of the Meta-Analytic Models.



Note. One effect size smaller than -10 and six effect sizes larger 10 not shown.

this request was rejected because the workload would be too high, we created another request targeting a shorter timespan https://web.archive.org/web/20240117112028/https://www.whatdotheyknow.com/request/trial_protocols_behavioural_insi_2 and https://www.whatdotheyknow.com/request/trial_protocols_behavioural_insi_2. This was also rejected with the justification that determining whether the department holds the information, locating, retrieving, and extracting it would take more than 3.5 days.

Figure 3.3: Aspects of Pre-Analysis Plans of Trials Run by Nudge Units.



for the 50 most recent pre-analysis plans, as of August 2022, and compared them to the final published reports. We did not analyse reports that did not include pre-analysis plans or did not include results (for example, because OES could not obtain the necessary data). We further skipped two trials that had conflicting registrations on OES and ClinicalTrials.gov, leaving us with a final sample of 32 reports with corresponding pre-analysis plans (see supplementary materials for details).

Our evaluation demonstrates several examples of best practices in OES pre-analysis plans. Some plans are highly detailed, including analysis scripts for later analyses (e.g., <https://web.archive.org/web/20240110143223/https://oes.gsa.gov/projects/soar/>). Additionally, 30/32 final reports at least detail the main outcome as described in the analysis plan, or otherwise disclose it transparently. There is, however, large variability in the quality of the pre-analysis plans and several plans have limitations. In particular, we find that both pre-analysis plans and final reports are usually insufficiently detailed to determine whether any selective reporting has taken place. 28/32 pre-analysis plans lack a sample size specification. This allows for ‘optional stopping’, where data is collected until a statistically significant result is found - a practice likely to inflate type 1 error rates (Simmons et al., 2011; Stefan & Schönbrodt, 2023). While OES often uses existing data, and therefore sample size justification may not be applicable, we noted that often the nature of the existing data (e.g., which agency will supply it or in which time period it will

be collected) was not clear. Further, the data are generally not publicly available. In many cases, this will be for sound legal reasons. However, making anonymized data available wherever possible is good practice as it allows other researchers to independently verify the claimed results and conduct additional robustness checks (Wagenmakers et al., 2022).

In 6/32 analysis plans, information about the covariate inclusion was lacking. This allows for analytic flexibility, where covariate specifications can be explored until a statistically significant result is found - again a practice that may inflate type 1 error rates (Simmons et al., 2011; Stefan & Schönbrodt, 2023). For example, one analysis plan¹⁴ determines demographic covariates to include in the regression for the main outcome as follows: ‘The precise way these demographic variables are categorized and included in the specification is not defined here.’ While this plan also includes a robustness check without covariate adjustment, the results of this check are not reported, and it is not possible to know which covariates are included for the test included in the final report. Indeed, 14/22 reports that pre-register robustness checks do not contain the outcomes of those checks and it is generally difficult to identify which covariates were included when estimating reported effect sizes. Finally, in 13/29 cases whose pre-analysis plans specify secondary analyses, these are not included in the final reports.¹⁵

Overall, those pre-analysis plans that have been shared are insufficient to rule out optional stopping or (intentional or unintentional) *p*-hacking. However, it is important to emphasize that this does not imply that, therefore, optional stopping or *p*-hacking has taken place - only that the existence of analysis plans does not strictly rule them out.

3.2.3 Evidence Based Public Policy Needs to Increase Transparency

We find that the pre-analysis plans and final reports lack sufficient detail to evaluate whether selective reporting has occurred, while statistical techniques provide

¹⁴<https://oes.gsa.gov/projects/transparent-defaults/>

¹⁵We also contacted the Office for Evaluation Sciences as well as DellaVigna and Linos to ask whether more detailed reports are available but received no response.

evidence for reporting biases. We call for more transparency, so that the quality of the work by nudge units can be independently evaluated by other researchers. Similar to recommendations for pre-analysis plans and transparency in academia (where similar problems have been identified; Brodeur et al., 2022; Claesen et al., 2021; G. Li et al., 2018; van den Akker et al., 2023), nudge units may increase transparency by taking several steps (roughly ordered by ease of adoption):

1. Analysis plans should be shared publicly by all nudge units.¹⁶ OES should be applauded for already doing so.
2. Analysis plans should be specific and include covariates for regression specification and either planned sample size or detailed information about the existing data set being used. In our supplements, we provide a recommendation for an updated OES analysis plan template that includes a specific section for sample size and treatment of covariates. In some cases, the dataset may not yet be shared with the nudge unit itself when the analysis plan is created. In these cases, it may not be possible to specify covariates in advance, or the anticipated covariates may be different from the ones available later. It is then important to be transparent about which covariates are included in the model and how they deviate from the analysis plan. Further, sensitivity analyses will then help to understand robustness to the covariate structure.
3. Write-ups of trial outcomes should be shared publicly and report the outcomes of all statistical tests that were specified in the pre-analysis plans. In general, more detail about the conducted analyses needs to be provided than is currently the case. If the main write-ups are intended to be short and for non-experts, another document with all analyses that were conducted should be shared (e.g., an R Markdown file).
4. The anonymised data and analysis code should be shared publicly. We are aware that this may not be possible in many cases (e.g., when medical records

¹⁶We are aware that sometimes contracts with clients may not allow doing so; however, we believe this is unlikely to be the case for all trials run (BIT), and recommend contracting in a way that allows sharing the pre-analysis plan, which is ultimately also in the interest of the clients.

are used); however, currently virtually no data are shared. We therefore urge BIT and OES to make the anonymized data available where this is legally possible.

5. Independent audits by third parties should take place to compare the pre-analysis plans against the reports. For example, behavioural insights teams could give small monetary awards to anyone who detects a mismatch between a published pre-analysis plan and corresponding report (similar to red team approaches that have been successfully applied in academia).¹⁷ Further, government agencies and other contractors should commission an evaluation of the work, when commissioning BIT or OES to run a trial (e.g., the UK Cabinet Office could fund PhD students to compare write-ups and pre-analysis plans). Note that it should be considered completely appropriate to deviate from an analysis plan or conduct additional analyses so long as deviations are transparent and justified and if confirmatory and exploratory analyses are clearly delineated in the report (Nosek et al., 2019).

One potential response to our suggestions is that BIT is a private company and thus should not be required by law to share pre-analysis plans or reports. While this may be a valid view, we point out that in most cases, BIT is in fact contracted by Government agencies. In these cases, where the taxpayer funds the research conducted, the contract should require the sharing of pre-analysis plans and of the outcomes of the research.

Further, we want to emphasize that nudge units have made an important contribution by popularizing RCTs within government. This allows researchers and policymakers to evaluate the effectiveness of different policy interventions and is an important pillar of evidence-based policy-making. We do not see our criticisms as showing the limitations of RCTs in general but only aim to point out specific and feasible improvements that nudge units could make to further enhance the effectiveness of their work.

¹⁷<https://web.archive.org/web/20240110143330/http://daniellakens.blogspot.com/2020/07/the-red-team-challenge-part-3-is-it.html>

There is great benefit in applying evidence-based behavioural science to public policy evaluated with randomised controlled trials and there are many examples of evidence from behavioural science positively affecting policy (Johnson & Goldstein, 2003). We also point out that OES has already taken several steps to increase transparency that go beyond many other government agencies. The inclusion of publicly accessible pre-analysis plans by all nudge units is a further step towards gold standards in behavioural science application. The evaluation we present is intended to motivate further strides towards fully transparent, evaluable, high quality research. We are confident that applied nudge units can embrace this challenge to the further benefit of society.

Reference for Journal Article: Maier, M.*, Bartoš, F.*, Raihani, N., Shanks, D.R., Stanley, T.D., Wagenmakers, E.-J., & Harris, A.J.L. (2024). Exploring open science practices in behavioural public policy research. *Royal Society Open Science*, *11*. <http://doi.org/10.1098/rsos.231486>.

Data and Code Availability: All data and code are available at <https://osf.io/f3rxt/>.

Supplemental Information: Additional results are provided in an electronic supplementary material at https://rs.figshare.com/collections/Supplementary_material_from_Exploring_open_science_practices_in_behavioural_public_policy_research_/7072542.

Author contributions: M.M.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, writing—original draft; F.B.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, visualization, writing—original draft; N.R.: conceptualization, investigation, supervision, validation, writing—review and editing; D.R.S.: conceptualization, investigation, supervision, validation, writing—review and editing; T.D.S.: conceptualization, investigation, methodology, supervision, validation, writing—review and editing; E.-J.W.: conceptualization, investigation, methodology, supervision, validation; A.J.L.H.: conceptualization, investigation, methodology, supervision, validation, writing—review and editing.

Part II

Experimental & Computational Modelling Work to Address Neglected Issues in Moral and Prosocial Decision Making

Background for Part 2

Part II of this thesis focuses on experimental work combined with computational modelling to investigate neglected issues in moral and prosocial decision making. The reason for pivoting away from publication bias adjusted meta-analysis and the topics in Part I is that I found many of the most important topics in moral and prosocial decision making had received too little attention in the previous literature (and were consequently not amenable to the application of the publication bias adjusted methods used in Part I). This includes topics such as how we can reduce scope insensitivity, how people make decisions under extinction risk (i.e., risks that, if materialized, could lead to human extinction), and how people learn to adopt certain moral beliefs. In contrast to these areas, the domains analysed in Part I already benefit from relatively well-established paradigms and methodologies, increasing the likelihood that others will take up replication efforts (as exemplified by recent large-scale projects, such as the multilab replication of Construal Level Theory: <https://climr.org/>).

Part II thus aims to broaden the scope of research on moral and prosocial decision making by shedding light on these overlooked areas and proposing new ways to study them. To do so, I extend existing paradigms (Chapter 4 on Scope Insensitivity & Unit Asking), and develop new paradigms with associated computational models (Chapters 5 & 6 on Decisions Under Extinction Risk and Moral Learning).

Chapter 4

Scope Insensitivity & (Sequential) Unit Asking

Everybody to count for one, nobody for more than one.

— Jeremy Bentham (as quoted in Mill, 1879, p.112)

Even though many people agree with the above premise of impartiality, in practice people usually do not help more as the size of a problem increases. This scope insensitivity is a well-studied limitation of human judgement, which describes the phenomenon that people do not scale their valuation of a quantity in proportion to its size or scope (Kahneman et al., 1999). It is commonly discussed in the context of contingent valuation studies, which investigate how people value different goods (Baron and Greene, 1996; Desvousges et al., 1993; Kahneman and Knetsch, 1992; Lopes and Kipperberg, 2020; Veisten et al., 2004) and charitable giving (Västfjäll & Slovic, 2020), which is the primary focus of the current research. Prior work, for example, has found that when participants were asked how much money they would donate to buy Christmas presents for 20 children, they donated no more than those asked to help a single child (Hsee et al., 2013). Scope insensitivity has also been well documented as a cause of neglecting human lives, explaining, for example, the relative indifference in response to some genocides (Cameron & Payne, 2011; Dickert et al., 2012, 2015; Slovic & Västfjäll, 2010a). The consequences of this phenomenon can therefore hardly be overestimated. While many papers discuss the

problem of scope insensitivity, a relatively minor proportion of the research output has been devoted to potential interventions to overcome it. A notable exception to this is Hsee et al.'s (2013) proposal of the Unit Asking technique.

4.1 Unit Asking

Unit Asking involves first asking participants how much they would be willing to donate to help one affected individual before extending to ask how much they would be willing to pay to help many people. According to Hsee et al. (2013), the underlying mechanism through which Unit Asking reduces scope insensitivity is people's desire for consistency (Ariely et al., 2003; Luisetti et al., 2011; Thomas et al., 2018). When people are first asked about one individual and in the next step asked about a larger number, their desire for consistency drives them to donate an amount that is proportional to the amount that they donated for one individual. Unit Asking has repeatedly been shown to be highly effective in increasing donations over 'Direct Asking' (where participants are solely asked about the total number of people and not initially asked about one individual; Karlsson et al., 2020; Marcinkiewicz, 2016; Simmons, 2013). For example, participants donated more than twice as much to help 20 children when they were first asked to think about how much they would donate to help one child, than when they were only asked to think about all 20 children ($M = \$49.42$ vs. $M = \$18.03$; Hsee et al., 2013).

4.2 Extending Unit Asking to *Sequential* Unit Asking

Given the applied relevance of increasing charitable donations and combating scope insensitivity, we aimed to improve upon the effectiveness of Unit Asking. In their original paper, Hsee et al. (2013) asked about one individual and then about a larger number of affected individuals. A natural step is, therefore, to extend this 'Classical' Unit Asking (henceforth CUA) by scaling up the scope sequentially (in a stepwise manner) to larger numbers of affected individuals. For example, when eliciting donations to fund Christmas gifts for 100 children, instead of asking about one child and then all 100, one could first ask about 1 child, then 2, and scale up by roughly doubling (e.g., 5, 10, 20, 50) until reaching 100. We expect that this Sequential

Unit Asking (SUA) will increase donations in comparison to CUA, as the repeated questions should provide additional ‘bite’ for a mechanism related to individuals’ desire for consistency to exert an influence. The first goal of this chapter is to test whether the SUA extension increases donations over CUA.

4.3 Does (Sequential) Unit Asking Make People Scope Sensitive?

We consider scope sensitivity a continuum, with complete scope insensitivity corresponding to a complete neglect of scope and maximal scope sensitivity reflecting linear proportionality (e.g., giving 100 times more to help 100 individuals than 1 individual). In between these two extremes are different magnitudes of scope sensitivity, such as logarithmic sensitivity (e.g., giving $\log(100)$ as much to 100 individuals as to 1 individual; Fechner, 1860), or an even weaker ordinal form of scope sensitivity, which only implies giving more to larger numbers of affected individuals (e.g., giving more to 100 individuals than to 1 individual). In this chapter, we focus on testing for this weak, ordinal form of scope sensitivity. Consequently, when we use the term ‘scope sensitivity’, it refers to ordinal sensitivity. If an asking technique does not even show this weakest form, it cannot show any of the stronger forms.

Whereas the increase in average donations through CUA has been well established (Hsee et al., 2013; Karlsson et al., 2020; Marcinkiewicz, 2016; Simmons, 2013), it is not yet clear to what degree this demonstrates scope sensitivity on the part of participants, even the weak ordinal form. For scope sensitivity, participants should give more as the number of affected people increases, even if the number of repeated questions remains the same (e.g., after being asked about one individual, they should donate more to help 10,000 affected individuals than to help 100). Hsee et al. (2013) recognise this in the discussion section of their paper. They cite unpublished data (Hsee, Zhang, & Lu, 2013) demonstrating that people gave more to help a total scope of 100 children than 10 children under CUA (and not Direct Asking).¹

¹We infer these simple effects from the fact that Hsee et al. (2013) report an interaction between

In contrast to Hsee et al. (2013), an unpublished Master's thesis (Marcinkiewicz, 2016) found CUA to only give a one-off boost to WTD judgements, independent of scope. Marcinkiewicz (2016) asked participants to donate to the charity *Global Alliance for Improved Nutrition* to help children affected by a food shortage in Mali. He used a CUA technique as described in the previous section - first asking participants how much they would be willing to donate to help one child and then how much they would be willing to donate to help 4, 20, 200, or 2000 children (this scope varied between-subjects). While participants donated more to help the group of children than a single child, the group donation did not differ between groups of 4, 20, 200, and 2000 affected children. This finding calls into question whether Unit Asking really increases scope sensitivity in willingness to donate (WTD) judgements or rather gives a single one-off boost. Consequently, more research is required to test whether CUA actually makes participants scope sensitive. Marcinkiewicz (2016) also showed that the higher the scope, the lower people's desire to be consistent. Because SUA presents a smaller increase in scope for each step, we predict that SUA will elicit scope sensitivity where CUA does not. The second goal of this chapter is to test whether either CUA or SUA can result in (ordinal) scope sensitivity, such that people give more to larger numbers of affected individuals.

In the following five studies, we first compare SUA, CUA, and Direct Asking (DA - *only* asking a single question about the full scope) and: a) replicate the beneficial effect of CUA over DA in WTD judgements; b) show that SUA increases WTD judgements over CUA (Studies 1a and 1b). In Study 2, we test whether CUA and SUA make participants scope sensitive by varying Asking Type and Scope. Study 2 failed to find evidence for scope sensitivity. Study 3 also observed no evidence for scope sensitivity, despite seeking to provide optimal conditions for it. Study 4 switches to contingent valuation instead of a WTD judgement dependent variable by asking participants how much money would be *needed* to help x individuals, rather

asking condition and scope, but no main effect of scope. Calculating the exact p -value of the interaction, $F(1, 310) = 4.15$, $p = .042$ shows that the claim about scope sensitivity was only weakly supported in Hsee et al. (2013). The underlying data cited as Hsee, Zhang, & Lu (2013) was - to our knowledge - also never published at a later point.

than how much they would *donate* to help x individuals. In the contingent valuation task, all asking techniques (including direct asking) showed scope sensitivity. This scope sensitivity was enhanced using SUA, but not CUA.

4.4 Study 1a

Study 1a was a non-preregistered pilot. In this study, we aimed to replicate the benefit of CUA in comparison to DA as in Hsee et al. (2013). In addition, we investigated whether the new SUA method could increase WTD judgements over those observed via CUA.

4.4.1 Method

Participants

This study received ethical approval from the University of Oxford Central University Research Ethics Committee (reference number R56657/RE001). Participants were paid \$0.24 for this 2-minute study. We uploaded the study on January 14th 2021 on Positly (<https://www.positly.com/>), selecting US participants as the target group. Positly is a front-end platform that recruits MTurk participants but adds additional quality metrics (<https://www.positly.com/participants/>). Positly blocks suspicious IPs, requires high approval rates, and requires participants to consistently pass attention checks. Initially, 406 participants signed up for the study. We excluded 3 participants that indicated that they were 1036, 1986, and 948 years old. After these exclusions, 403 participants remained, of which 194 were female, 207 were male, and 2 indicated another gender. The mean age was 38.27 (SD = 11.61).²

Design

Participants were randomly assigned to one of three experimental conditions: Direct Asking (DA), Classical Unit Asking (CUA), Sequential Unit Asking (SUA). The key dependent variable was the amount (in USD) participants were willing to donate

²Because of the extreme brevity of the task, we did not add an attention check in this study. However, we added an attention check in Study 3, finding that it led to no additional exclusions.

to help 100 children, entered as a free response.

Materials and Procedure

Following Hsee et al. (2013), participants were first asked to imagine that the principal of a neighbourhood kindergarten had sent them an email to ask for money that would be used to buy Christmas gifts for children:

Imagine the following: Christmas is around the corner. The principal of a neighborhood kindergarten has sent you an email asking for donations. You know her personally and trust her words. The email directs you to a website with the following questions. Please answer these questions as if you were making actual donation decisions.

They were then directed to a site that described a kindergarten with 100 children from low-income families and asked participants how much they would be willing to donate:

Thanks for visiting our website. Please read the following carefully and answer the ensuing questions. Even if you are not willing to make a donation, please still answer the questions; you may simply enter \$0. You can revise your answers, and your answers will not be recorded until you move on to the next page.

Our kindergarten currently has 100 children (like the one pictured below), they are all from low-income families. And their parents have little money to buy Christmas gifts for them. We hope you can make a donation, so we can buy Christmas gifts for them.

We used 100 children instead of the 20 used in the original study, since a setup with a larger scope seemed more appropriate to test the sequential asking technique. The experiment had three conditions.

In the control condition (as in Hsee et al., 2013), participants were asked directly how much they would be willing to donate to buy Christmas gifts for 100 children:

Please think about all **100** of these children. How much are you willing to donate to help these 100 children? Please enter the amount of money you decide and agree to donate: __ \$

In the CUA condition (as in Hsee et al., 2013), participants first indicated how much they would be willing to donate to help one child and only afterwards (on a separate page) asked how much they would be willing to help 100 children. In other words, we first asked:

Before you decide how much to donate to help these 100 children, please first think about **one** such child and answer a hypothetical question: How much would you donate to help this one child? Please indicate the amount here: __ \$'

After filling in this amount, they were asked the following question on the next page:

Now please think about all **100** of these children. How much are you willing to donate to help these 100 children? Please enter the amount of money you decide and agree to donate: __ \$

Finally, in the SUA condition (novel to this study), participants were also asked to indicate how much they would be willing to donate to help one child first (as in the CUA condition). However, instead of extending directly to 100 children, we scaled up the scope sequentially by asking (on separate pages) their WTD judgments for 2, 5, 10, 20, 50, and only then 100 children. In general, the increase per step can be determined by the n th root of the full scope, where n is the number of steps (i.e. here, $\sqrt[6]{100}$). This increase is founded in Fechner's law, postulating that the subjective intensity of stimulus corresponds logarithmically to the stimulus intensity (Fechner, 1860). This law suggests that increasing the number of children by a constant multiplier will correspond to increasing by a constant sum in terms of participants' subjective stimulus intensity. We additionally round to the next multiple of five for small numbers and the next multiple of ten for larger numbers to make the numbers more intuitive to participants (e.g., in this study, we use 50 rather than 46 as implied by the formula).

Participants provided their responses by typing any amount they saw fit (in US dollars) into a free text response box.

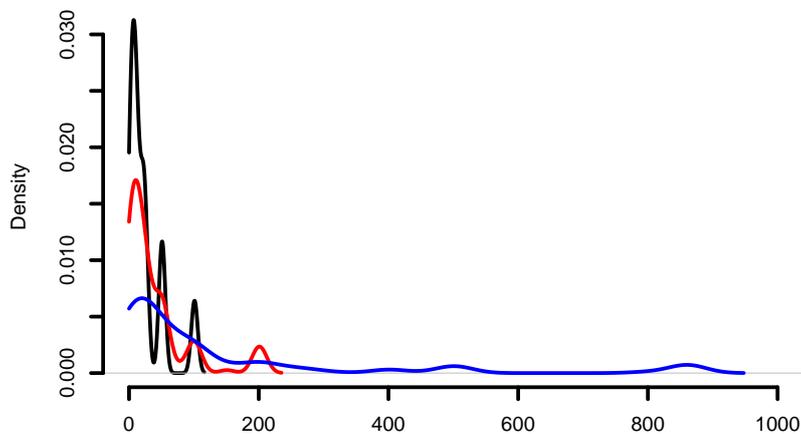
Analyses

Difficulties in Analysing WTD Judgements. Open response WTD judgements often result in large outliers due to the lack of an upper bound. Previous papers on Unit Asking address this by *winsorizing* outliers and then employing a *t*-test or ANOVA (Hsee et al., 2013; Karlsson et al., 2020). Winsorizing replaces values above the 95th percentile of values provided with the value exactly at the 95th percentile, to reduce the skew in the data. However, winsorizing is often not effective as even after winsorizing, the data is extremely skewed. This can make inferences overly dependent on the small subset of participants that indicate extremely large WTDs. We illustrate this using the data of Study 1a as an example. Figure 4.1 shows the distribution of WTD judgements in the three different groups *after* winsorizing. The distributions are still extremely skewed, and hence the application of methods assuming normally distributed residuals is not appropriate. Winsorizing is also theoretically dubious. If a participant indicates a very high WTD judgement, it seems plausible that: (1) they did not take the task seriously and should be excluded; and/or (2) that they would actually donate that much. We are not aware of any mechanism which would give rise to the idea that they meant to indicate the value that the 95th percentile participant indicated (winsorizing). We can also consider the influence of individual datapoints on the observed pattern of significance. As an example of how strongly outliers affect statistical significance patterns under winsorizing, let us consider the comparison of DA (black) and CUA (red) groups. Naively analysing the winsorized data indicates that CUA is more effective than DA, $t(202.95) = 3.0677$, $p = .002$. However, even though the comparison contains 271 participants and the p -value is small, the pattern of significance hinges upon the WTD judgements of 6 participants. If we remove the 6 highest values in the CUA group, the difference between the groups no longer attains the accepted significance level, $t(226.26) = 1.916$, $p = .057$. This should give some intuition about

how unstable the classical inference using winsorized data is in WTD judgements.

Our Analytic Approach. Due to the highlighted limitations associated with winsorizing, we take a different approach in this chapter, which provides Bayesian models with lognormal likelihood, and non-parametric frequentist solutions to accommodate the skew in observed responses. Because we cannot meaningfully analyse scope insensitivity for participants who donated \$0 (as they would donate \$0 for any scope under scope consistency as well as scope inconsistency), and because the lognormal model cannot accommodate zeros, we excluded these participants in the following analyses. Our approach evolved during the project and this was the pre-registered approach for Study 3. This is the analysis we focus on for all studies in this chapter. An alternative write-up with analyses exactly as preregistered is available in the OSF project. The conclusions are the same in terms of the Bayes factor categories unless stated otherwise in footnotes.

Figure 4.1: Densities of WTD Judgements in Study 1a After Winsorizing.



Note. Black is the DA group, red the CUA group, and blue the SUA group.

To test the differences between conditions, we specified a Bayesian analysis using the package BRMS (Bürkner, 2017, 2018). We always compared the likelihood of the data under one model assuming a difference in WTD judgements to a model assuming no difference in WTD judgements using the `bridgesampling`

package (Gronau, Sarafoglou, et al., 2017; Gronau, Singmann, & Wagenmakers, 2017). We refer to comparisons of two groups as *t*-tests and to comparisons of more than two groups as ANOVAs. Because our data was expected to be positive and quite skewed, we used a lognormal likelihood rather than the more common Gaussian likelihood ($WTD_i \sim \text{Lognormal}(\mu_i, \sigma^2)$, where *i* denotes the condition). As well as visually confirming the shape of the data, we also validated the likelihood function by comparing lognormal and normal likelihood models with leave-one-out cross validation (Vehtari et al., 2017), which indicated superior performance of the lognormal models. In addition, the exponent of μ should usually approximate the empirical median of the data on a linear scale, although the two can differ where the data cluster around prominent numbers (e.g., 20, 50). This is observed in our data. We present posterior medians (i.e., $\exp(\mu)$) in addition to the empirical medians in our visualization of the results.

The analysis was conditional on participants donating at all (i.e., by removing zeros), and we also removed values larger than \$10,000 (and note any instances where inferences are affected by this exclusion). We checked that our intervention did not affect the number of people donating in the first place, using a Bayesian contingency table analysis in the BayesFactor R package (Morey et al., 2015), with assumed sampling type joint multinomial and prior concentration one.

For the main analysis, we used a prior of $\text{normal}(\mu = 3, \sigma = 1)$ on the intercept. We used a prior of $\text{normal}(0.4, 0.4)$ on the main effect of Asking Type. Finally, we used a prior of $\text{normal}(1, 0.5)$ on σ . As the median of the lognormal distribution corresponds to $\exp(\mu)$ this implies that we expect a median donation of $\exp(3) = 20.9$ in the control group and $\exp(3 + 0.4) = 29.96$ in the intervention group. In addition, these priors result in reasonable prior predictives for this kind of donation task (see Appendix B.3, i.e., the mean and CI predicted from the priors are similar to values we would expect).

We use Bayes factors as our primary inference criteria (Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995; Rouder & Morey, 2019; Wrinch & Jeffreys, 1921). A Bayes factor compares the probability of the data under the null

(no effect) to the alternative as specified by the prior distributions outlined above. As a general rule of thumb, Bayes factors between 1 and 3 are regarded as anecdotal evidence, Bayes factors between 3 and 10 are regarded as moderate evidence, and Bayes factors larger than 10 are regarded as strong evidence (Jeffreys, 1939; M. D. Lee & Wagenmakers, 2013). The inverse of the Bayes factor can be used to describe evidence for the null hypothesis. For example, a Bayes factor between 1/3 and 1/10 would be considered moderate evidence for the null. For robustness, we also analysed the data with frequentist non-parametric tests.

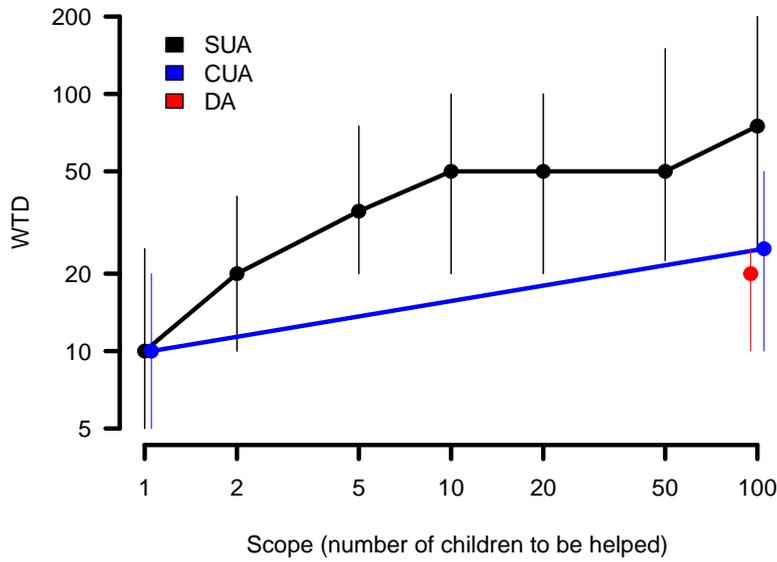
4.4.2 Results and Discussion

Effect on Proportion of Participants Donating and Summary Statistics

A Bayesian contingency table analysis indicated that the intervention did not affect the share of participants donating in the first place (DA: 10.37%, CUA: 13.43%, SUA: 14.18%; $BF_{10} = 0.032$; see Appendix B.4 for the share of participants donating at all across conditions in all experiments). Our main analysis excludes WTD judgements of 0 (51 participants) and larger than 10,000 (zero participants).

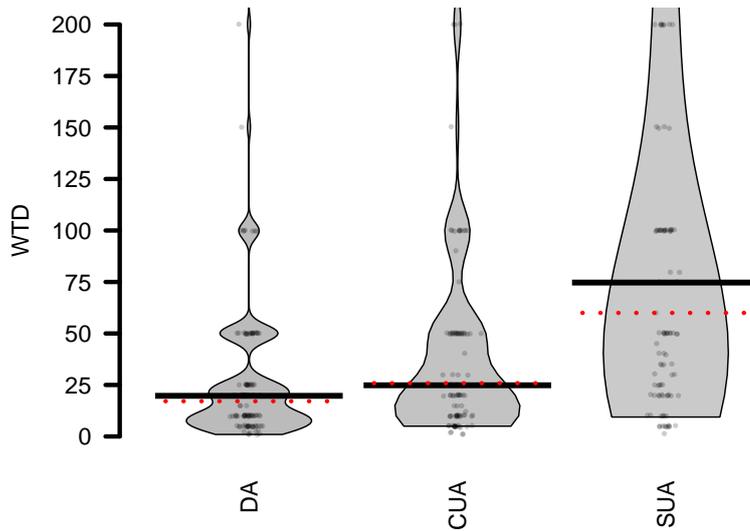
Figure 4.2 visualizes the median WTD judgements for all WTD questions asked (i.e., per step), whereas Figure 4.3 visualizes the WTD judgements only for the total scope of 100 children. The median donation for 100 children in the DA condition was \$20, the median donation for 100 children in the CUA condition was \$25, and the median donation for 100 children in the SUA condition was \$75. For one child, the median donation was \$10 in both intervention conditions. A Wilcoxon test shows no evidence that SUA and CUA conditions differed in terms of donations for one child ($W = 6526$, $p = .774$).

Figure 4.2: Median WTDs for Each Step in Study 1a.



Note. Error bars indicate the interquartile range.

Figure 4.3: WTD Distribution for the Full Scope of 100 in Study 1a.



Note. Grey area indicates empirical density. Black lines indicate empirical medians. Red lines indicate posterior medians (i.e., the median of the posterior distribution for this parameter).

Effect of SUA and CUA on Donations

The 3-level ANOVA on WTD judgements to 100 children indicated overwhelming evidence for an effect of condition ($BF_{10} = 4.65 \times 10^9$). Pairwise comparisons indicated strong evidence for an effect of CUA in comparison to the DA condition ($BF_{10} = 33.08$), overwhelming evidence for SUA in comparison to the DA condition ($BF_{10} = 22.43 \times 10^{10}$), and strong evidence for SUA in comparison to CUA ($BF_{10} = 22101.89$). The results were also corroborated with a frequentist analysis using non-parametric Wilcoxon tests (all p -values $< .01$). As predicted, this study: (1) replicated the effectiveness of CUA in increasing WTD judgements compared to DA; (2) showed that our extension to SUA gives a considerable additional boost over CUA.

4.5 Study 1b

4.5.1 Method

This study was a preregistered direct replication of Study 1 (<https://osf.io/uchq8>).

Participants

This study received ethical approval from Harvard University's ethics review board. We paid participants \$0.31 for this 2-minute study. The study was uploaded on February 3rd 2021 on Positly for US participants. Initially, 507 participants signed up for the study. We excluded one participant who indicated that their age as 3963. The mean age was 40.69 ($sd = 12.64$). 248 participants were female, 256 were male, one indicated gender 'other', and one did not indicate their gender.

4.5.2 Results and Discussion

Effect on Proportion of Participants Donating and Summary Statistics

A Bayesian contingency table analysis indicated that the intervention did not affect the share of participants donating in the first place (DA: 11.76%, CUA: 13.43%,

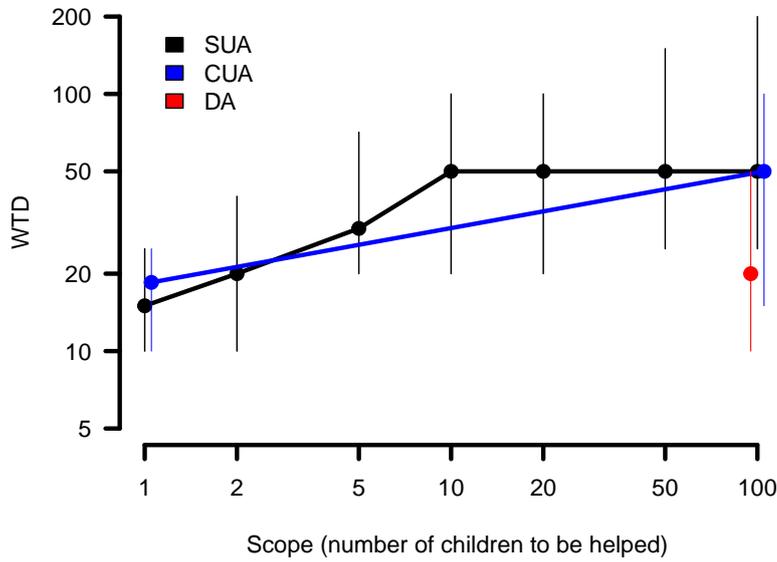
SUA: 14.18%; $BF_{10} = 0.015$). We excluded WTD judgements of 0 (60 participants) and larger than 10,000 (3 participants; conclusions are unaffected by this exclusion unless indicated in a footnote). We proceeded to the main analysis with the remaining 443 participants. Figure 4.4 displays WTD judgements per step after these exclusions. Figure 4.5 shows the distribution of WTDs for the full Scope of 100 children. The median donation for 100 children in the DA condition was \$20, the median donation for 100 children in the CUA condition was \$50, and the median donation for 100 children in the SUA condition was \$50. Note that if many participants indicated exactly the median, the distribution of WTD judgements between groups may still differ even though the medians are the same. Figure 4.5 shows that this is indeed the case and more WTD judgements are above the median for SUA, compared with CUA, which is also reflected in the posterior medians (red line). The median donation to help one child was \$18.50 in the CUA condition and \$15 in the SUA condition. A Wilcoxon test shows no evidence that SUA and CUA conditions differ in terms of donations for one child ($W = 11267$, $p = .453$).

Effect of SUA and CUA on Donations

The 3-level ANOVA indicated overwhelming evidence for an effect ($BF_{10} = 2.63 \times 10^{11}$).³ Pairwise comparisons indicated moderate evidence for an effect of CUA in comparison to the DA condition ($BF_{10} = 22.04$), strong evidence for SUA in comparison to the DA condition ($BF_{10} = 18.46 \times 10^8$), and strong evidence for SUA in comparison to CUA ($BF_{10} = 5\,643.68$). The results are corroborated with a frequentist analysis using non-parametric Wilcoxon tests (all p -values $< .01$).

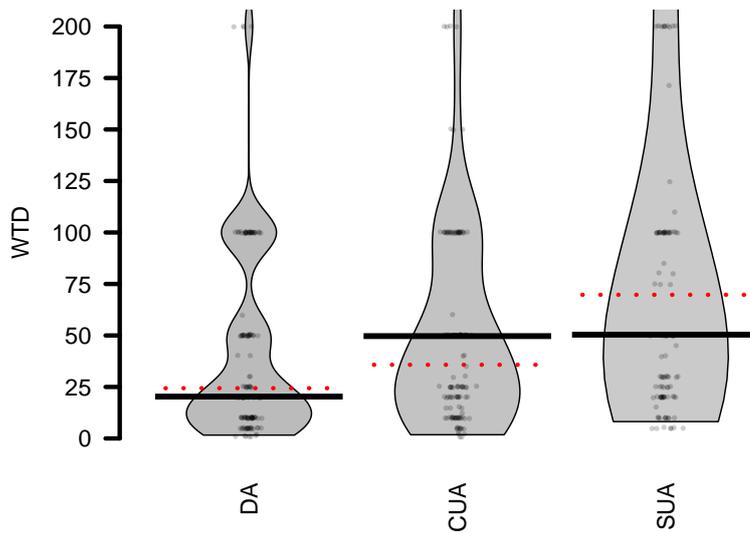
³Only 3.08 when not excluding the participants donating more than \$10,000. The change in evidence is quite large as one of the excluded participants indicated they would donate 1 million dollars. As this is likely not indicative of their actual behaviour, we believe more weight should be given to the analysis with exclusions.

Figure 4.4: Median WTDs for Each Step in Study 1b.



Note. Error bars indicate the interquartile range.

Figure 4.5: WTD Distribution for the Full Scope in Study 1b.



Note. Grey area indicates empirical density. Black lines indicate empirical medians. Red lines indicate posterior medians (i.e., the median of the posterior distribution for this parameter).

In summary, Study 1b replicated the result that CUA increased WTD judgements over DA and that SUA additionally increased WTD judgements in comparison to CUA.

4.6 Study 2

Studies 1a and 1b showed that SUA elicited higher average WTD judgements for groups of 100 than DA and CUA. However, to demonstrate scope sensitivity, we need to show that the donation is larger when more units are under consideration (e.g., people donate more to help 10,000 people than to help 100 people). Therefore, Study 2 not only varied Asking Type but also the maximum Scope (100 vs. 10,000).

When scaling up SUA to a higher scope, there are fundamentally two possible approaches. First, one can keep the increase per step constant. In this case, a larger number of steps will be required to still reach the same full scope for a larger scope size (e.g., 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10,000). Second, one can keep the number of steps constant. This will result in a higher increase per step, which is required to reach a larger scope with the same number of steps (e.g., 1, 5, 20, 100, 500, 2000, 10,000). In Study 2, we investigated both of these options.

Hypotheses

We preregistered three hypotheses:

1. In a 2×2 analysis with factors Asking Type (Direct vs. CUA) and Scope (100 vs. 10,000), there will be no interaction of Asking Type and Scope (replicating Marcinkiewicz, 2016).
2. In a 2×2 analysis with factors Asking Type (CUA vs. SUA) and Scope (100 vs. 10,000), there will be an interaction of Asking Type and Scope such that SUA minus CUA is positive, and larger for Scope 10,000 than Scope 100.
3. If the effect of SUA is partially driven by an increase in the number of steps, we will observe that participants in the increase per step constant condition

(SUAI) would donate more for 10,000 children than those in the number of steps constant condition (SUA).⁴

4.6.1 Method

Participants

This study received ethical approval from the Ethics Chair for the Department of Experimental Psychology, UCL (Project ID No: EP/2021/001). We paid participants between \$0.40 and \$0.65 for participating in a 3-5 minute study (depending on condition). The study was uploaded on August 24th and 25th, 2021, targeting US participants via Positly. Based on our preregistered stopping rule, we used Bayesian sequential analysis starting with 280 participants, adding 140 additional participants in two steps until we reached 560 participants (including WTD judgments of 0). Due to multiple participants taking part in the Study in parallel, 562 participants signed up, of which 285 were female, 275 male, and 2 indicated gender ‘other’, with mean age of 40.16 (sd = 12.07).

Experimental Conditions

We employed a similar method to Study 1 with a number of changes. To allow more realistic scaling up to larger scopes, we replaced the kindergarten example with a vignette describing a food shortage in Ethiopia. Participants were told that in one area of the country, 100 or 10,000 children were affected by the shortage and asked how much they wanted to donate to the charity ‘Global Alliance for Improved Nutrition’ to help the children in need (based on Marcinkiewicz, 2016). The combinations of Scope and Asking Type required to answer our Research Questions resulted in seven conditions. Participants were randomly assigned into one of the following conditions (with the exact sequence of the number of children mentioned stated in parentheses):

⁴We initially preregistered to also look at an ANOVA interaction in this case; however, we realized that for Scope 100, the two Asking Types do not differ. Therefore, we used a *t*-test only on Scope 10,000.

1. Scope 100 x DA (100)
2. Scope 10,000 x DA (10,000)
3. Scope 100 x CUA (1, 100)
4. Scope 10,000 x CUA (1, 10,000)
5. Scope 100 x SUA (1, 2, 5, 10, 20, 50, 100)
6. Scope 10,000 x SUA - number of steps constant (1, 5, 20, 100, 500, 2000, 10,000)
7. Scope 10,000 x SUAI - increase per step constant (1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10,000)

Each Unit Asking condition also included the following, exploratory, desire for consistency measure: ‘You said you would give X dollars for a single child and Y dollars for the group of N children. In doing so, were you trying to be consistent? That is, were you trying to allocate for each child in the group of N children as much as money as for the single child? (1 = not at all, 4 = somewhat 7= yes, absolutely)’. The variables X, Y, and N were adapted based on the participant’s responses. The consistency measure was asked at the end of the study to avoid any influences on the key variable of interest (WTD).

Analyses

We used the same analysis as Study 1b with the addition of a prior of normal(0.25, 0.25) on the interaction. The reason why the priors are increasingly more narrow is that, given the lognormal likelihood, the prediction on the linear scale corresponds to the exponent of the marginal mean. Therefore, using similar priors on main effects and interactions would result in the prediction of an extremely large interaction effect on the linear scale. We tested for the presence of the hypothesized interactions by comparing models that include the interaction to models that do not include the interaction using Bayes factors. In addition, we conducted a frequentist analysis

using rank-based (non-parametric) tests for main effects and interactions with the package `rfit` (Kloke, McKean, et al., 2012), and independent sample Wilcoxon tests when comparing pairwise differences.

4.6.2 Results

As we did not achieve sufficient evidence on all three critical tests in the preregistered intermittent analyses, we collected the full sample consistent with our preregistration.

Effect on Proportion of Participants Donating and Summary Statistics

A Bayesian contingency table analysis again indicated that the intervention did not affect the share of participants donating more than zero ($BF_{10} = 0.19$; see Table B.1 for proportions). Therefore, we proceeded to the main analysis excluding WTD judgements of 0 (133 participants) and larger than 10,000 (exclusions by condition: DA = 2; CUA = 2; SUA - number of steps constant = 11; SUA - increase per step constant = 4; conclusions are unaffected by this exclusion unless indicated in a footnote), leaving us with a total of 410 participants.

Figure 4.6 shows the median WTD judgements per step and Figure 4.7 shows the distribution of WTD judgements for the total number of children (see also Table B.1). A Wilcoxon test shows no evidence that SUA and CUA conditions differ in terms of donations for one child ($W = 10992$, $p = .767$).⁵

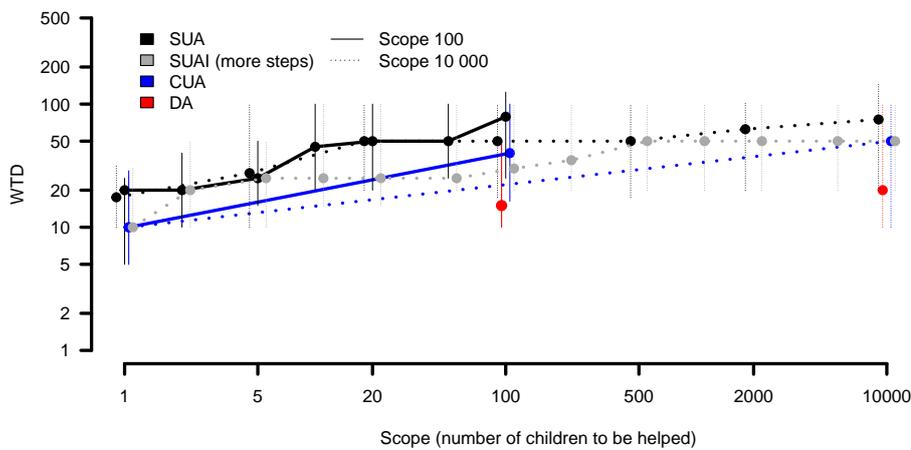
H1: Comparing CUA and DA

The first hypothesis, that CUA would not make participants scope sensitive, received only weak support. A Bayesian interaction test in a 2x2 ANOVA with factors Scope (100 vs. 10,000) and Asking Type (DA vs. CUA) implied weak evidence

⁵The reason why the medians in the Figure for one child are different even though there is no significant difference is that participants cluster their responses around prominent numbers (i.e., 10, 20, 50). Therefore, only a few participants changing their judgement can result in a jump between two prominent numbers.

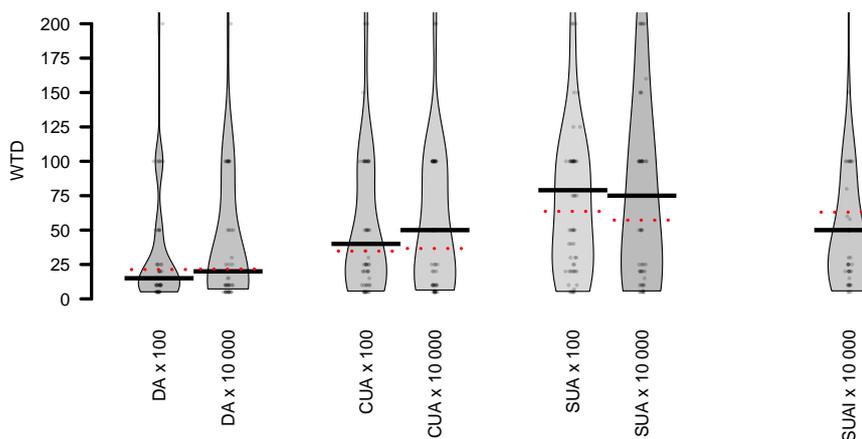
against an interaction of Asking Type and Scope ($BF_{10} = 0.50$).⁶ In addition, there was strong evidence for a main effect of CUA vs. DA ($BF_{10} = 10.03$)⁷. A non-parametric ANOVA also found no evidence for an interaction of Asking Type and Scope, $F(1, 228) = 0.49$, $p = .488$, and a significant main effect of CUA vs. DA, $F(1, 228) = 5.11$, $p = .025$.

Figure 4.6: Median WTDs for Each Step in Study 2.



Note. Error bars indicate interquartile range.

Figure 4.7: WTD Distribution for the Full Scope in Study 2.



Note. Grey area indicates empirical density. Black lines indicate empirical medians. Red lines indicate posterior medians. SUAI is for SUA with increase per step constant.

⁶We found strong evidence under the preregistered model.

⁷We found only moderate evidence when not excluding WTDs larger than \$10,000 ($BF_{10} = 6.48$).

H2: Comparing SUA and CUA

We found weak evidence against the second prediction that SUA would result in more scope consistency than CUA, as revealed by weak evidence against an interaction between SUA (number of steps constant) vs. CUA and Scope ($BF_{10} = 0.64$).⁸ As in the previous study, we found a main effect of SUA vs. CUA ($BF_{10} = 11.70$),⁹ although the evidence here is much weaker. These findings were corroborated with a non-parametric, frequentist analysis finding no interaction, $F(1, 247) = 0.35$, $p = .552$, and a (just) significant main effect of SUA vs. CUA; $F(1, 247) = 3.96$, $p = .047$. Further, when only comparing the increase in WTD judgements under SUA vs. CUA for Scope 100, which is similar to Study 1, the evidence for the effect is only moderate ($BF_{10} = 4.085$).¹⁰ This is surprising given the strong evidence found in Study 1.

H3: Comparing Different Types of SUA

We found no evidence that donations were higher for 10,000 children in the SUAI (increase per step constant) condition than the SUA (constant number of steps) condition, as revealed by a t -test ($BF_{10} = 0.67$).¹¹ This is also supported by a non-parametric Wilcoxon test, $W = 1473$, $p = .949$.

Exploratory Analyses: Investigating Judgements in Intermediate Steps and Desire For Consistency

The donation patterns so far give rise to some interesting avenues for more detailed examination. Two conditions used identical sequences of steps up to 100 participant: SUA with constant increase per step for Scope 10,000, dotted grey line in Figure 4.6; SUA for Scope 100, solid black line in Figure 4.6. This raises the question

⁸We find moderate evidence for this interaction when not excluding participants donating more than \$10,000 ($BF_{10} = 3.61$), and strong evidence against this when using the preregistered model

⁹Moderate evidence in the preregistered analysis

¹⁰Only 2.37 without the exclusions.

¹¹Moderate evidence that they do not differ under the preregistered analysis.

Table 4.1: Share of Participants Showing a Strong Monotonic Increase in Scope 10,000 SUAI (Increase per Step Constant) Condition and Scope 100 SUA Condition.

Step	SUA - maximum Scope 100	SUAI - maximum Scope 10,000
1 → 2	52%	43%
2 → 5	53%	42%
5 → 10	45%	34%
10 → 20	33%	24%
20 → 50	36%	19%
50 → 100	31%	18%
100 → 200		19%
200 → 500		20%
500 → 1000		19%
1000 → 2000		11%
2000 → 5000		15%
5000 → 10,000		15%

of why the WTD judgements are not higher for the full scope in SUAI with constant increase per step for Scope 10,000, given the larger number of affected individuals as well as larger number of steps. Two explanations come to mind: (1) participants reach some kind of donation ceiling after which they do not donate more; (2) participants (who were informed in the beginning what the maximum scope would be) think ahead and donating less for one child when they know the full scope will be higher. In other words, they might first form a judgement about what to donate according to the full scope and then donate a proportional fraction of this to one child. Table 1 shows the share of participants that indicate a strong monotonic increase at each step (i.e., donate more for more children). Table 1 confirms what is suggested by Figure 4.6. In the Scope 100 condition, more participants increase on each step in comparison to the Scope 10,000 condition. In addition, the proportion of participants increasing is lower for the larger scopes in the Scope 10,000 condition. In other words, both looking ahead and reaching a donation ceiling may play a role in this donation behaviour.

Which kinds of participants are more likely to keep increasing their donations, and which are more likely to drop out? One hypothesis is that participants that donate more for one child are more likely to drop out later, as they are more likely

to run out of money. To test this, we conducted a linear regression predicting the number of strong monotonic increases from the WTD judgements for a single child. However, we did not find any evidence for the notion that WTD judgements for one child were related to monotonic increases for any of the 3 SUA conditions (maximum Scope 100: $F(1, 79) = 2.18, p = .501$, maximum Scope 10,000 & increase per step constant: $F(1, 77) = .764, p = .385$, and maximum Scope 10,000 & number of steps constant: $F(1, 77) = .666, p = .417$).

Finally, we found evidence against an effect of Asking Type on our exploratory measure, 'desire for consistency' ($BF_{10} = 0.055$), which we included in line with Marcinkiewicz (2016).

4.6.3 Discussion

We again replicated the benefit of SUA over CUA; however, the evidence was much weaker than in Study 1 (though still moderate in strength).

We found no evidence that unit asking increased scope sensitivity in this study. This finding holds for SUA and CUA - both techniques only give a constant boost independent of scope. In other words, repeated asking leads people to indicate higher WTD judgements, but this effect appears *independent* of Scope (the number of children affected). Perhaps the most surprising result is that we do not observe scope sensitivity in the SUAI - increase per step constant - condition. To explain this result, we suggest that some participants might reach a donation ceiling (as suggested by the smaller number of participants displaying monotonic donation patterns for the larger number of affected individuals - see Table 4.1). Further, some participants might simply find the large number of repetitive questions that were asked in this condition unpleasant or irritating and, therefore, disengage from the task. Study 3 included alternative optimal conditions for observing scope sensitivity on the basis of these conjectures.

4.7 Study 3

Study 3 aimed to provide the most favourable conditions for scope sensitivity. To reduce the likelihood of participants reaching a donation ceiling, we reduced the

maximum scope in this study from 10,000 to 50. Achieving scope consistency for such a smaller scope is likely more realistic, and Hsee et al. (2013, p.1806) would appear to agree: ‘We should also note, however, that the ability of unit asking to increase scope sensitivity is likely limited; if the target numbers are large — for example, 1,000 versus 10,000 — respondents may encode either number as ‘a lot’ and not differentiate the two’. To avoid participants planning forward when knowing the maximum scope, we included conditions where we do not tell participants the maximum scope in advance of the iterative procedure. Finally, we also reverted back to the kindergarten vignette that we had used in Study 1 which showed the strongest benefit of SUA over CUA. Overall, we aimed to test the following four preregistered research questions in this study:

1. Can we identify scope sensitivity for any of the asking techniques?
2. Does SUA increase WTD judgements over CUA?
3. Does SUA make participants more scope sensitive than CUA?
4. Does telling people the maximum scope beforehand affect scope sensitivity?

4.7.1 Method

Participants

This study received ethical approval from the Ethics Chair for the Department of Experimental Psychology, UCL (Project ID No: EP/2021/001) and was preregistered at <https://osf.io/ezrs9>. We paid participants \$0.37 for participating in a 2-minute study. The study was uploaded between the fourth and sixth of November 2021, targeting US participants via Positly. Initially, 574 participants signed up, no participants were excluded by age, and none failed the attention check. The mean age was 40.31 (sd = 12.62). 289 participants were female, 279 were male, and 6 reported ‘other’.

Experimental Conditions

To test the research questions, we used several combinations of Asking Type and Scope. Unless otherwise indicated, and unlike in Studies 1 and 2, participants were *not* informed of the maximum scope in advance. Participants were randomly assigned into one of the following conditions (with the exact sequence of steps stated in parentheses):

1. Scope 10 x CUA (1, 10)
2. Scope 50 x CUA (1, 50)
3. Scope 10 x SUA (1, 2, 5, 10)
4. Scope 50 x SUA – number of steps constant (1, 4, 15, 50)
5. Scope 10 x SUA (1, 2, 5, 10) [participants know maximum scope when answering first question]
6. Scope 50 x SUA – number of steps constant (1, 4, 15, 50) [participants know maximum scope when answering first question]
7. Scope 50 x SUAI – increase per step constant (1, 2, 5, 10, 20, 50)

We did not include a DA condition this time as, at this point, it is well established that people are scope insensitive under DA, and that both CUA and SUA increase WTD judgements relative to this baseline.

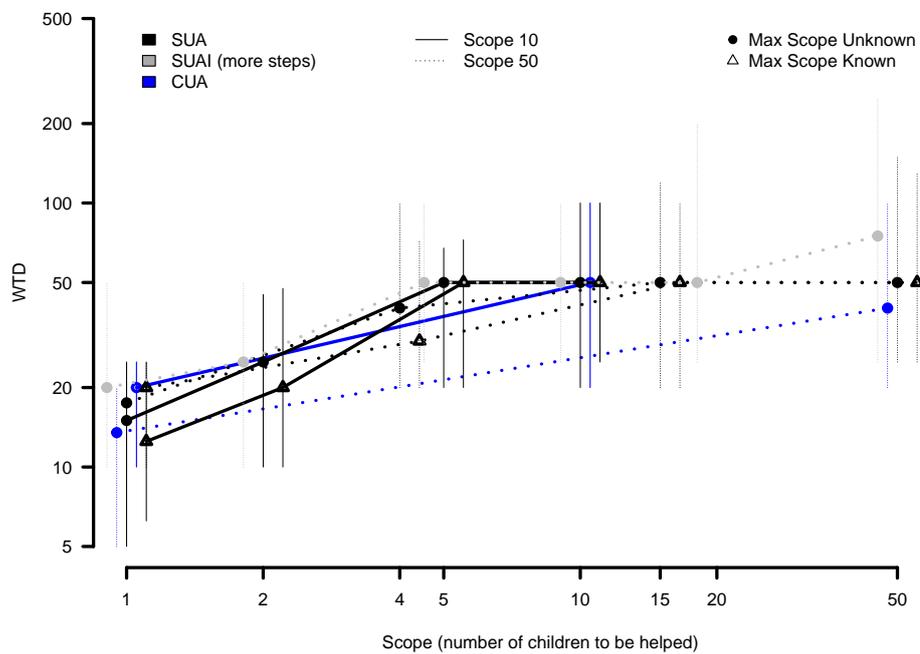
4.7.2 Results

Effect on Proportion of Participants Donating and Summary Statistics

We used the same statistical model as in Study 2. Asking Type did not affect the number of participants donating in the first instance ($BF_{10} = 0.03$, see Table B.2 for proportions). We proceeded to our main analysis, excluding participants that

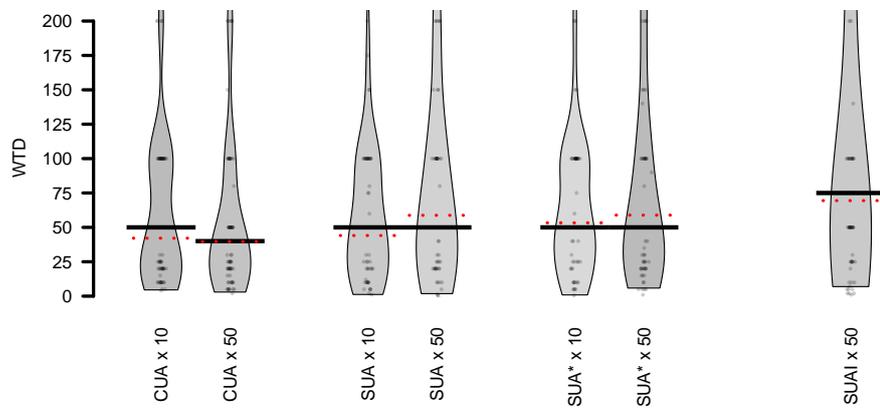
donated 0 (70 participants) and participants that donated more than 10,000 (0 participants). Figure 4.8 visualizes the median donation trajectories for the different steps, and Figure 4.9 shows the distributions of WTD judgements for the full scope (see also Table B.2). In line with the random assignment, a Wilcoxon test shows no evidence that SUA and CUA conditions differed in terms of donations for one child ($W = 26393, p = .916$).

Figure 4.8: Median WTDs for Each Step in Study 3.



Note. Error Bars Represent the Interquartile Range.

Figure 4.9: WTD Distribution for the Full Scope in Study 3.



Note. Grey area indicates empirical density. Black lines indicate empirical medians. Red lines indicate posterior medians (i.e., the median of the posterior distribution for this parameter). SUA I is for SUA with increase per step constant.

*SUA conditions where the maximum Scope was known to participants. In all other conditions the maximum Scope was unknown.

RQ1: Can We Identify Scope Sensitivity in Any Format?

We tested this with four comparisons between the WTD judgements for 50 vs. 10 participants for both CUA and SUA (i.e., in terms of the Experimental Conditions 1 vs. 2, 3 vs. 4, 3 vs. 7, and 5 vs. 6). We found moderate evidence that participants donate the same for 50 vs. 10 individuals for CUA ($BF_{10} = 0.29$). For SUA, we compared the Scope 10 condition to the Scope 50 condition (number of steps constant) and the Scope 50 condition (increase per step constant). We found no evidence for either the null or alternative hypothesis for the number of steps constant comparison ($BF_{10} = 0.86$). For the increase per step constant comparison, we found weak evidence that participants donated more with larger scopes ($BF_{10} = 2.55$). In addition, for SUA (number of steps constant), we also have one condition where participants knew the maximum scope when starting the experiment. Here we find weak evidence against participants donating more for larger scopes ($BF_{10} = 0.36$). In sum, Study 3 did not reveal convincing evidence for scope sensitivity in any format. This lack of evidence is also corroborated by non-parametric Wilcoxon

tests. This indicates no evidence for scope consistency under CUA ($p = .357$), no evidence for scope consistency under SUA (number of steps constant, $p = .241$), no evidence for scope consistency under SUAI (increase per step constant, $p = .077$), and no evidence for scope consistency when telling people the maximum scope beforehand ($p = .883$).

RQ2: Does SUA Increase WTD Judgements Over CUA?

Comparing SUA (number of steps constant) to CUA (1 & 2 vs. 3 & 4), we find no evidence for or against an effect of SUA ($BF_{10} = 0.99$).¹² The Wilcoxon test is (just) significant in favor of SUA ($W = 9159$, $p = .047$). When instead comparing SUAI (increase per step constant, 1 & 2 vs 3 & 7) to CUA we find weak evidence for a difference: Bayesian analysis: $BF_{10} = 2.68$, Frequentist analysis: $W = 9919.5$, $p = .018$. Note, however, that we preregistered to use SUA (number of steps constant) to test RQ2.

RQ3: Does SUA Make Participants More Scope Sensitive Than CUA?

For the interaction between Asking Type (SUA [number of steps constant] vs. CUA) and Scope (10 vs. 50; i.e., 1 & 2 vs 3 & 4), we find no evidence that SUA increases scope consistency in comparison to CUA ($BF_{10} = 1.17$). When instead testing this interaction using SUAI (increase per step constant; i.e., 1 & 2 vs. 3 & 7), we find weak evidence that SUAI increases scope consistency in comparison to CUA ($BF_{10} = 2.48$). This finding is partially in line with a non-parametric interaction test, which shows no evidence for an interaction in the number of steps constant condition, $F(1, 287) = 3.06$, $p = .081$; however, it does show evidence for a significant interaction for SUAI (increase per step constant), $F(1, 286) = 5.21$, $p = .023$.

¹²Using only SUA (number of steps constant)

RQ4: Does Telling People the Maximum Scope Beforehand Affect Scope Sensitivity?

When testing an interaction between telling people the maximum scope before and Scope (i.e., 3 & 4 vs. 5 & 6), we find no evidence that telling people the maximum scope beforehand increases scope sensitivity ($BF_{10} = 0.42$) for the SUA conditions. This is also confirmed by a robust frequentist ANOVA, $F(1, 281) = 0$.

Additional Analyses

Does the Effect of SUA vs. CUA differ between Studies 2 and 3? In this study, we did not find evidence for the main effect of SUA. Importantly, we also did not find evidence *against* an effect of SUA. To investigate whether this result is driven by the smaller scope in this study, we compared the SUA vs. CUA effect in this study (for the condition with 50 participants) to that in Study 1a (Scope = 100), where we found the strongest effect of SUA. We tested the interaction between Condition (SUA vs. CUA) and Study to investigate whether the difference in conditions was affected by the study. As we did not have a Unit Asking condition where the scope was known in advance in Study 3, we used the conditions where the scope is not known for this comparison. We only found weak (and non-significant) evidence for this interaction (Bayesian analysis: $BF_{10} = 2.32$; nonparametric ANOVA: $F(1, 373) = 1.400$, $p = 0.237$). We conclude, therefore, that the effect of SUA vs CUA does not reliably differ across the studies.

Effect of SUA vs. CUA Across All Donation Studies. To further test the overall effect of SUA, we pooled the data from all four studies and one study presented in Appendix B.1, which employed a different design that likely diminished the effectiveness of SUA. We tested for an overall effect of SUA vs. CUA across these five studies using a Bayesian mixed-effects model with random effects for Study and Scope. We used the same priors as in Study 3 and BRMs' default priors on the random effects.¹³ When pooling across studies, we only find moderate evidence for

¹³ $t(3, 0, 2.5)$ on random intercepts and slopes and LKJ(1) on the correlation between random

an effect of SUA in comparison to CUA ($BF_{10} = 3.65$). If we only look at the studies in the main text of this chapter, the evidence is somewhat stronger ($BF_{10} = 4.50$). In conclusion, when pooling across all relevant studies, there is moderate evidence that SUA results in higher WTD judgements for helping multiple individuals than does CUA.

4.7.3 Discussion Study 3

We again did not observe any evidence for scope sensitivity, even after creating these most favorable conditions to observe scope sensitivity in donation judgements. We believe the two most likely explanations for this result are:

1. Even with reduced scope, participants' budget constraints limit their donation judgements to an extent that scope sensitivity may not be observed with WTD judgements.
2. None of the asking techniques promotes scope sensitivity in general.

If the first explanation is correct, we would expect to see evidence for scope sensitivity in a contingent valuation version of the study. In other words, instead of asking participants how much money they would donate to buy Christmas gifts for the children, one would ask how much money they think is required to buy Christmas presents for these children. As no willingness for a personal donation is asked for, budget constraints would no longer explain the lack of scope sensitivity. We tested such a contingent valuation setup for Study 4.

4.8 Study 4

To switch to a contingent valuation setup in Study 4, we replaced the question about donations in Study 3 with the following question:

Please think about all 50 [10 in the lower scope condition] of these children.
How much money do you think is needed to buy Christmas gifts for these 50
children? Please indicate the amount here: ..\$

effects

We modified the background scenario from Study 3 by telling participants that they are active in the local community and occasionally give advice to people in their community when they have administrative questions. In a next step, we informed them that the principal of a neighbourhood kindergarten has contacted them to ask how much money they think is needed to buy Christmas gifts for the children. We elicited the contingent valuation judgements with the different asking techniques specified in Study 3.

The aim of this study was to test the following preregistered research questions:

1. Does Classical Unit Asking (CUA) increase judgements over Direct Asking (DA)?
2. Does Sequential Unit Asking (SUA) increase judgements over CUA?
3. Which asking techniques induce scope sensitivity?
4. Does CUA make participants more scope sensitive than DA?
5. Does SUA make participants more scope sensitive than CUA?

4.8.1 Method

Participants

This study was approved by the Ethics Chair for the Department of Experimental Psychology, UCL (Project ID No: EP/2022/001) and preregistered at <https://osf.io/abcd12>. Participants were paid \$0.46 for their participation in a 2.5-minute study. The study was uploaded on the 10th of February, 2023 via Positly. Initially, 567 participants signed up, 1 participant was excluded because they did not indicate their age, and 28 failed the attention check (a question asking how many children were in the kindergarten). The mean age of participants was 41.26 (SD = 11.77) with 268 being female, 265 male, and 5 reporting 'other'.

Experimental Conditions

To test the preregistered research questions, we randomly assigned participants to one of the following combinations of Asking Type and Scope:

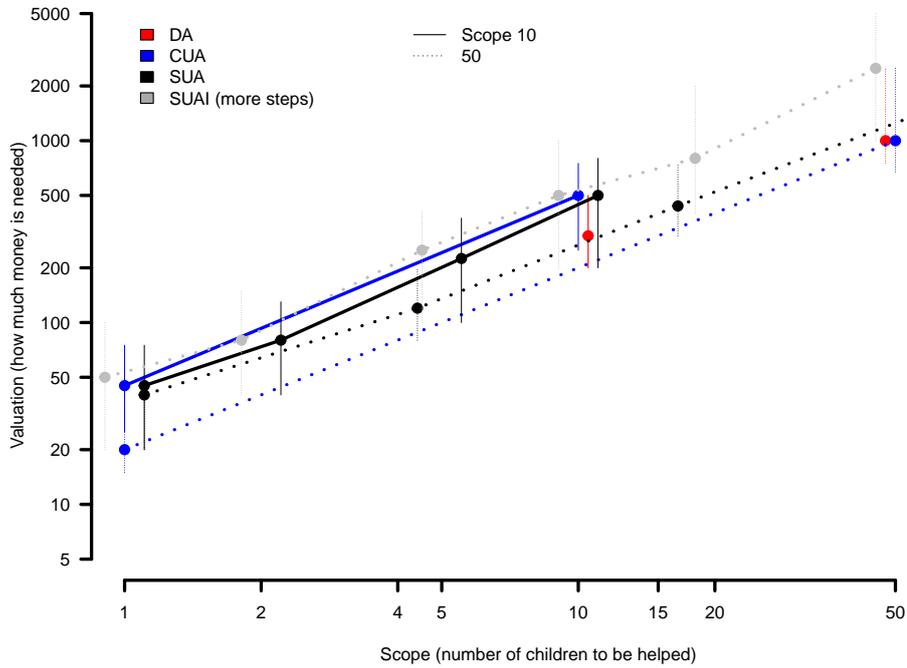
1. Scope 10 x DA (10)
2. Scope 50 x DA (50)
3. Scope 10 x CUA (1, 10)
4. Scope 50 x CUA (1,50)
5. Scope 10 x SUA (1, 2, 5, 10)
6. Scope 50 x SUA – number of steps constant (1, 4, 15, 50)
7. Scope 50 x SUAI – increase per step constant (1, 2, 5, 10, 20, 50)

4.8.2 Results and Discussion

Effect on Proportion of Participants Donating and Summary Statistics

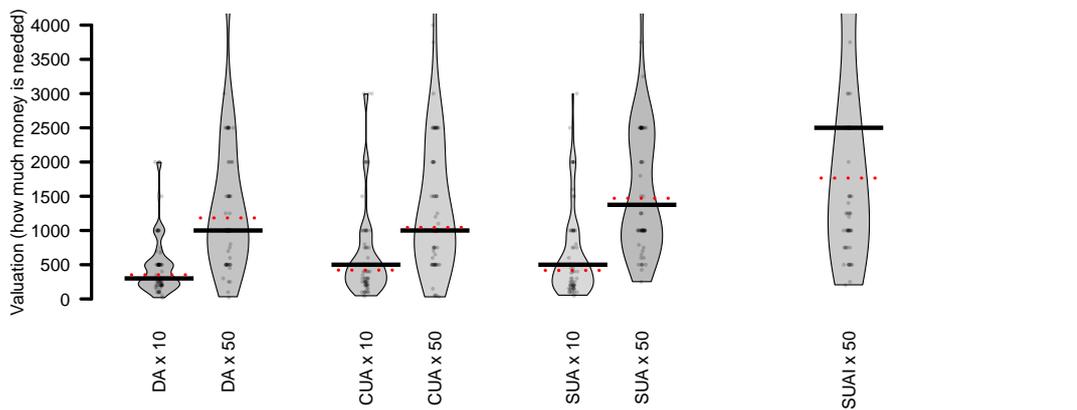
The statistical model was the same as in Study 3. However, this time none of the participants indicated an amount of zero, thus, we did not conduct the contingency table analysis and proceeded directly to our main analysis, excluding participants who indicated more than 10,000 (12 participants, four for each asking type; conclusions are unaffected by this exclusion unless indicated in a footnote). Figure 4.10 visualizes the median contingent valuation trajectories for the different steps, and Figure 4.11 shows the distribution of contingent valuation judgements for the full scope. The figures suggest that this time participants did increase their judgements for larger numbers of affected children. In line with the random assignment, a Wilcoxon test shows no evidence that SUA and CUA conditions differed in terms of valuations for one child ($W = 15195$, $p = .140$).

Figure 4.10: Median Contingency Valuation Judgements for Each Step in Study 4.



Note. Error Bars Represent the Interquartile Range.

Figure 4.11: Contingency Valuation Distribution for the Full Scope in Study 4.



Note. Grey area indicates empirical density. Black lines indicate empirical medians. Red lines indicate posterior medians (i.e., the median of the posterior distribution for this parameter). SUAI is for SUAI (increase per step constant).

RQ1 & RQ2: Which Asking Technique Induces the Highest Valuation Judgements?

When comparing DA and CUA (Conditions 1 & 2 vs 3 & 4), we find no evidence for a difference in judgements (Bayesian analysis: $BF_{10} = 0.21$, frequentist analysis: $W = 11060$, $p = .653$).

When comparing SUA to CUA, we also find no evidence for a difference in judgements when using the number of steps constant condition (Conditions 3 & 4 vs 5 & 6; Bayesian analysis: $BF_{10} = 0.87$, frequentist analysis: $W = 10378$, $p = 0.329$); however, we find moderate evidence with the increase per step constant, SUAI, condition (Conditions 3 & 4 vs 5 & 7, Bayesian analysis: $BF_{10} = 6.13^{14}$, frequentist analysis: $W = 1093$, $p = .102$).

RQ3: Which Asking Techniques Induce Scope Sensitivity?

When comparing donations for Scope 10 and Scope 50 we find strong evidence for scope sensitivity for all of the asking techniques:

- **DA** (Conditions 1 vs 2): Bayesian analysis: $BF_{10} = 3.46 \times 10^{12}$; frequentist analysis: $W = 843$, $p < .001$.
- **CUA** (Conditions 3 vs 4): Bayesian analysis: $BF_{10} = 2.48 \times 10^6$; frequentist analysis: $W = 1186.5$, $p < .001$.
- **SUA** (Conditions 5 vs 6): Bayesian analysis: $BF_{10} = 1.59 \times 10^{15}$; frequentist analysis: $W = 771.5$, $p < .001$
- **SUA** (increase per step constant; Conditions 5 vs 7): Bayesian analysis: $BF_{10} = 1.82 \times 10^{16}$; frequentist analysis: $W = 724$, $p < .001$

¹⁴Only weak evidence when not excluding donations larger than 10 000, $BF_{10} = 2.54$

RQ4: Does CUA Make Participants More Scope Sensitive Than DA?

We find evidence against the hypothesis that CUA makes participants more scope sensitive than DA (interaction of Conditions 1 & 2 vs 3 & 4; Bayesian analysis: $BF_{10} = 0.32$, frequentist analysis: $F(1, 298) = 0.33$, $p = 0.564$).

RQ5: Does SUA Make Participants More Scope Sensitive Than CUA?

For the number of steps constant condition, we find moderate evidence that SUA makes participants more scope sensitive than CUA (interaction of Conditions 3 & 4 vs 5 & 6; Bayesian analysis $BF_{10} = 4.16$, frequentist analysis, $F(1, 294) = 4.31$, $p = .039$)¹⁵, while for the increase per step constant condition we find strong evidence for this effect (interaction of Conditions 3 & 4 vs 5 & 7; Bayesian analysis: $BF_{10} = 36.05$, frequentist analysis: $F(1, 297) = 12.915$, $p < .001$)

Conclusion Study 4

For contingent valuation judgements, we only find evidence for a benefit of SUA over CUA in average judgements when using the increase per step constant condition, SUAI, rather than the number of steps constant condition. We further find evidence for an interaction between Scope and CUA vs SUA, for both SUA types. In other words, SUA may only increase contingent valuation judgements if: (1) the scope is larger than ten and (2) a sufficient number of steps is asked (in this case five steps). Further, we find that people are scope sensitive in a contingent valuation setting even under DA. Notably, this scope sensitivity can further be enhanced using SUA but not using CUA.

¹⁵Without the exclusion of donations larger 10,000 $BF_{10} = 2.93$, frequentist analysis, $F(1, 294) = 2.63$, $p = .106$

4.9 General Discussion

We tested whether Unit Asking makes people scope sensitive as claimed in previous research. We found that Unit Asking only gives a one-off boost to WTD judgements, independent of scope. In other words, participants donated more under Unit Asking as opposed to Direct Asking; however, this increase was independent of the number of individuals affected and, therefore, does not seem to reflect genuine scope sensitivity.

We also introduced a new variant of Unit Asking, which we call Sequential Unit Asking (SUA). SUA extends Classical Unit Asking (CUA) by asking a sequence of questions scaling up with scope. We found evidence in three out of four studies that SUA increased WTD judgements over CUA. In addition, when pooling across all studies we found overall evidence that SUA increased WTD judgements over CUA. However, this increase also seems to come only as a limited series of one-off boosts rather than covarying with scope.

We further investigated contingent valuation judgements, where we found the inverse pattern in comparison to WTD judgements. People were scope sensitive under all of the asking techniques but CUA and SUA did not strongly increase judgements in comparison to direct asking. Table 4.2 gives an overview of the results across all five studies.

Finally, we advanced the methodology for analysing willingness to donate (WTD) judgements. We showed that the oft-used method of winsorizing and using conventional t -tests is problematic, as even after winsorizing the inference is not robust to outliers. Instead, we used a Bayesian model with lognormal likelihoods that can directly accommodate skew. We share the analysis code in our OSF project so that other researchers might use the methodology when analysing WTD judgements, or other judgements with similar distributional properties (i.e., extreme positive skew & zeros are reasonably excluded).

Table 4.2: Summary of Results Across All Four Studies

Study	Mean (Median) DA	Mean (Median) CUA	Mean (Median) SUA	Evidence for SUA over CUA	Evidence for Scope Sensitivity CUA	Evidence for Scope Sensitivity SUA
Donation Judgements						
Study 1a	\$27.36 (\$10)	\$81.13 (\$20)	\$177.58 (\$50)	+++	N/A	N/A
Study 1b	\$42.14 (\$20)	\$72.70 (\$25)	\$231.01 (\$50)	+++	N/A	N/A
Study 2	\$85.92 (\$10)	\$136.74 (\$25)	\$257.01 (\$50)	+++	–	–
Study 3	N/A	\$86.36 (\$30)	\$97.27 (\$50)	–	–	–
Contingent Valuation Judgements						
Study 4 (Scope 10)	\$479.34 (\$300)	\$637.97 (\$500)	\$627.77 (\$500)	–	+++	+++*
Study 4 (Scope 50)	\$1853.51 (\$1000)	\$1697.73 (\$1000)	\$2059.93 (\$1375)			

Note. ‘+++’ denotes Strong evidence for \mathcal{H}_1 , ‘++’ denotes moderate evidence for \mathcal{H}_1 , ‘+’ denotes weak evidence for \mathcal{H}_1 , ‘–’ denotes weak evidence for \mathcal{H}_0 , ‘– –’ denotes moderate evidence for \mathcal{H}_0 , and ‘– – –’ denotes strong evidence for \mathcal{H}_0 . Medians and Means here reported including donations of 0. SUA uses the number of steps constant condition for scaling up the scope. For Study 2, SUA refers to the condition, where the maximum scope was not known for comparability. As only for Study 4 the judgements differed strongly based on scope, we distinguish between Scope 10 and Scope 50 for this study.

* Unlike CUA, SUA further increased scope sensitivity above direct asking.

4.9.1 Why is it Difficult to Create Scope Sensitivity in WTD Judgements and Easier in Contingent Valuation?

A possible explanation for the difficulty of making people scope sensitive in WTD judgements using the different Unit Asking manipulations is based on mental accounting. In line with research on this topic (Sussman et al., 2015; Thaler, 1985, 1999), participants might have a fixed budget of how much they are willing to give to charity. This allocated money might already be exploited on relatively small numbers of children; therefore, participants do not usually increase their WTD judgements anymore for larger scopes in the Unit Asking condition. This is also reflected in comments that we got from participants in Study 2, where we included an open feedback box. For example, one participant indicated ‘I have a certain amount total I’m willing to give (\$15), so after that is reached I’m not willing to give anymore.’

and another participant said ‘I have limited funds so I can only donate so much regardless of the number in need.’ Versions of these comments were echoed by a considerable proportion of participants. Even though we tried to mitigate the problem by reducing the maximum scope in Study 3, it is still possible that both studies had reached the donation ceiling for most participants on both scopes, thus diminishing scope sensitivity.

In contrast, the contingent valuation setting removes this possibility: participants are asked how much money is *needed* to buy Christmas presents. This explains why participants are more readily scope sensitive even without using CUA or SUA. This baseline scope sensitivity can be further enhanced by SUA. However, we only observe a weak form of scope sensitivity, where participants donate more for more children, but this number does not increase proportionally with the scope, a finding that is qualitatively in line with previous literature on scope insensitivity (Dickert et al., 2012, 2015).

Overall, we conclude that when (hypothetical) personal funds are at stake, participants are reluctant to even show weak scope sensitivity, and only strong manipulations such as repeated asking will induce increases in donations. On the other hand, people readily show scope sensitivity for contingent valuation judgements.

4.9.2 Future Directions

In this chapter, we have focused solely on hypothetical WTD judgements. A standard suggestion for future research would therefore be to include incentive compatible studies. We do not, however, perceive these as a fruitful avenue for future research in this area for three reasons. First, in the current studies, our observations of different results between WTD judgements and contingent valuation judgements, coupled with participants’ references to budgetary constraints in open-text feedback, suggests our participants did take the hypothetical judgements seriously, and were restricted by real-world budgets. Second, the long ongoing debate about how much financial incentives change participant behaviour in the social and behavioural sciences (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001), mostly concludes that ‘the effects of incentives are mixed and complicated’ (Camerer & Hogarth,

1999, p. 1). While research has found that reliance on hypothetical donations increases mean donations (Bekkers, 2017), to our knowledge there is no evidence that hypotheticality influences differences between conditions in WTD judgements. Finally, incentive compatible studies typically provide an endowment and ask participants to donate a proportion of this (e.g., Schoenegger & Costa-Gomes, 2022; Small et al., 2007), which is a considerably different task from the majority of real-life donation decisions. Given that our participants appeared to take the WTD task seriously (as evidenced above), we would see such an incentive compatible task as more different from real donation decisions than our hypothetical tasks, especially as it induces an upper bound on how much participants can donate (the endowment), which would make it more difficult to study scope sensitivity.

While we do not see the hypotheticality of the current judgements as limiting the generalisability of the current results, we suggest future research should seek to generalise them beyond an experimental setting. For real donation decisions, there may be a trade off between the ‘boost’ provided by SUA and the need to maintain potential donors’ attention and reduce the number of questions asked. This question could not be answered within any lab / experimental context. Regardless of whether incentives are offered or not, experimental participants have a reasonable expectation of answering a series of questions. To determine how to maintain potential donors’ interest and lever the benefits of SUA in the real-world, a naturalistic field study would be required. Such a study would enable stronger conclusions as to the effectiveness of SUA in increasing charitable donations ‘in the wild.’

Further, in line with previous studies on Unit Asking (Hsee et al., 2013; Karlsson et al., 2020), we only include a picture of one needy child and don’t increase the number of children in the picture as scope increases. The fact that the Unit Asking method increases average donations provides evidence that participants are able to scale their concern with the number of children, at least in an ordinal way, even when visualizations of the scope are not provided. However, adding visualizations to represent additional children may help induce scope sensitivity in combination with (Sequential) Unit Asking. Identifying appropriate visualisations

might therefore be one avenue for future research to further increase scope sensitivity (see, for example, this educational video for visualization/animation techniques <https://www.youtube.com/watch?v=LEENEFaVUzU>).

Finally, future work may also explore different sequences of steps. Study 2 showed that the effect of SUA levels off when including too many steps. It would be interesting to further investigate at which point the effect of SUA begins to level off (although leveling off would likely depend on a number of contextual factors, such as the maximum scope). In addition to changing the number of steps, one could also change the ordering of steps, or include steps in random order to further investigate how the effects are shaped by the step order.

4.9.3 Conclusion

This chapter showed the difficulty in inducing scope sensitivity in WTD judgements. Contrary to previous claims, Unit Asking does not increase scope sensitivity in separate evaluation, but instead gives a one-off boost to WTD judgements. However, our findings also suggest that this one-off boost is increased by asking additional stepwise questions, a technique we term Sequential Unit Asking.

Reference for Journal Article: Maier, M., Caviola, L., Schubert, S., & Harris, A.J.L. (2023). Investigating (Sequential) Unit Asking – An Unsuccessful Quest for Scope Sensitivity in WTD Judgements. *Journal of Behavioural Decision Making*, 36(4), e2335. <https://doi.org/10.1002/bdm.2335>.

Data, Code, and Preregistration Availability: All data, code, and preregistrations are available at <https://osf.io/qjuh/>.

Author contributions: All authors were involved in the conceptualization and design of the experiments. M.M. created the materials and programmed the experiments with feedback from the other authors. M.M. conducted data analysis. M.M. wrote the first draft of the manuscript. All authors contributed to writing, reviewing, and editing the manuscript.

Chapter 5

Decisions Under Extinction Risk

In the previous chapter, we documented the pervasive influence of scope insensitivity on human decision making. This chapter turns to a key neglected area, where scope insensitivity may distort human decision making: decisions under extinction risk. In everyday life, people routinely make decisions that involve irredeemable risks such as death (e.g., while driving). These decisions under extinction risk are common, practically important, and have different properties compared to the types of decisions typically studied by decision scientists; for instance, the total number of choices that can still be made has a considerable impact on the optimal strategy (and to the extent that people are scope insensitive they may fail to take differences in this time horizon into account). Despite their practical importance and theoretical relevance, decisions under extinction risk have received little research attention. The first part of this chapter introduces a new task to study decisions under extinction risk in individual decisions, while the second part demonstrates how the paradigm can be generalised to decision making about collective risks (e.g., extreme climate change). Following one of my key recommendations for mitigating publication bias in Part I of this thesis, I submitted the collective task as a Stage 1 registered report. Therefore, the section on collective decision making under extinction risk only includes a pilot study, which shows the feasibility of the approach and demonstrates that collectives are more risk-taking than individuals.

5.1 Individual Decision Making Under Extinction

Risk

Throughout their daily lives, people make decisions involving risks that could change their lives as they know them: a potential jaywalker decides whether to cross the street in front of an incoming car, a driver decides whether they have enough space to make an overtaking manoeuvre, and a boxer decides whether to risk a career-ending injury in a high-stakes bout. At a collective level, people need to decide how much to invest in managing small probability, high-stakes risks, such as extreme climate change or pandemics.

The common denominator across all these examples is the risk of irredeemable extreme outcomes. Hopefully without sounding too ominous, we refer to such examples as *decisions under extinction risk*. Extant literature, recent history, and everyday anecdotes (*‘that driver’*) suggest that people are bad at making decisions about these types of risks (Elga et al., 2024; Wiener, 2016). However, in any real-world decision under extinction risk, a large set of psychological factors are at play simultaneously and it is difficult to disentangle the relative influence of different factors. For instance, inadequate social distancing during the Covid-19 pandemic could have been the result of self-interested decision making (especially with regards to younger people for whom COVID-19 was on average less severe), scope insensitivity (i.e., people do not take the large number of people affected from a pandemic outbreak into account), or underestimating the probability and cost of getting severely ill or even dying from the disease, to name just a few of many plausible factors.

Anyone tackling the challenges posed by extinction events will find little relief in the pages of mainstream decision making research. To see this, consider decision making research’s ‘fruit-fly’, the binary-choice lottery paradigm (e.g., Kahneman & Tversky, 1979; Rieskamp, 2008). In this paradigm, participants are presented with one lottery problem at a time, for example choosing between a sure win of £50 versus a 50% chance of winning £110. These lottery problems are independent

in the fullest sense of the term: at no point do outcomes of these decisions affect the possibility of future choices or their outcomes. By contrast, the outcomes of decisions under extinction risk might reduce the options available in future decisions – or prevent any future decision making at all.

Of course, the binary lottery paradigm is only one among the many different options available to researchers. Paradigms such as the Balloon Analogue Risk Task (Lejuez et al., 2002, 2003; Pleskac, 2008; Pleskac et al., 2008; Wallsten et al., 2005), or the Bomb Risk Elicitation Task (BRET, Crosetto and Filippin, 2013), mitigate some of the limitations of the binary lottery paradigm. However, these tasks were not designed to study extinction risks as discussed in this chapter. They therefore come with several limitations that make them unsuitable to study decisions under extinction risk, such as dependencies between choice and event probabilities, impossibility of learning from experience, or the probability of the extinction event being much higher than realistic in real-world settings (for a detailed review of potentially relevant tasks and their limitations, see Appendix C.1).

Given that existing paradigms do not allow studying extinction risks, what predictions can be derived from the extant decision making literature? Prospect theory would suggest that low probability events are overweighted and people are loss-averse (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). We might therefore expect people to be risk-averse for decisions involving low-probability, high-impact options (and thus avoid jaywalking). However, this overweighting of low probabilities does not necessarily extend to contexts in which people learn about the probabilities from experience, a discrepancy known as the the description-experience gap (Hertwig et al., 2004; Kellen et al., 2016; Lejarraga and Hertwig, 2021; Wulff et al., 2018). For many extinction risks, people often lack access to descriptive information and instead rely on personal experience (or the experiences of those they know). Further complicating this picture, other studies involving experienced options point to people overweighting extreme events (i.e., events with extremely good or bad utilities; see Lieder, Griffiths, and Hsu, 2018; Lieder et al., 2015; Ludvig et al., 2014; Madan et al., 2014; Sunstein and Zeckhauser, 2011). As

extinction events clearly have extreme (dis)utility, this literature would predict their overweighting.

In addition to the probability of the extinction event, another factor that would likely influence people's decisions is the perceived badness of extinction. Several well-documented psychological effects imply an underestimation of the value lost by extinction. For instance, scope insensitivity (i.e., not valuing a good in proportion with its scope or size; Desvousges et al., 1993) and narrow bracketing (i.e., viewing one choice in isolation rather than as part of a wider set of choices; Read et al., 2000) may cause people to underestimate the cost of being precluded from a larger number of choices or experiences in the future. Likewise, opportunity cost neglect (Frederick et al., 2009) may cause people to underestimate the opportunity cost of not being able to do what they could have done had they avoided an extinction event.

These factors would influence decision making even if people had accurate, unbiased information about the probability of the extinction event. These factors are likely amplified by biases in the available samples, in particular survivorship or observer selection effects (Ćirković et al., 2010; Wilson, 2023): conditional on thinking about the probability of the extinction event, the decision maker has not been subjected to one themselves, and will therefore have experienced zero instances of it (e.g., someone who is reflecting on the risk of jaywalking will have zero personal experience with deadly accidents, as otherwise they would not be in a position to ponder this question). The impossibility of sampling extinction events from memory will likely lead to underestimation of extinction risk.

In summary, several considerations point toward risk aversion, such as the overweighting of small probabilities from descriptions or of extreme events, while others suggest risk-seeking tendencies, including the description–experience gap, underestimation of extinction's utility loss, and observer selection effects. Given the conflicting predictions that can be drawn from past research and the practical importance of the topic, it is surprising that there is a 'notable scarcity of research on human behavior in the context of low-probability high-impact events' (Sundh, 2024, p.7). One of the main reasons for this absence may be the lack of a be-

havioural paradigm to study these types of risks, especially one that can be connected with normative benchmarks.

The present work contributes to the understanding of how people make choices under extinction risk. After providing the necessary clarification on what is meant by an *extinction event*, and offering two distinct definitions, we introduce an experimental paradigm suitable for its study – the *Extinction Gambling Task* (EGT). The EGT, like most decision making paradigms, isolates key features of decisions under extinction-risk within a lab-based framework involving financial gains and losses (Frey et al., 2017; Kahneman & Tversky, 1979). This approach not only enables a controlled first exploration of such decisions but also supports the derivation of optimal choice strategies for different variants of extinction events, which can serve as performance benchmarks.

After introducing the EGT, we report three experiments demonstrating its utility in examining people’s choices under extinction risk. Experiment 1 shows that participants are sensitive to differences between two different kinds of extinction events in ways that qualitatively align with optimal strategies. Experiment 2 introduces a new type of catastrophic non-extinction event and shows that participants treat this type of event differently from the extinction events, again consistent with optimal predictions. Finally, Experiments 3a and 3b demonstrate how the EGT can be used to study factors that might plausibly influence people’s decision making under extinction risk.

5.1.1 Towards an Extinction Gambling Task (EGT)

It is typical for scientific investigations to call for regimentations of ordinary-language concepts through the development of technical definitions (e.g., going from *warmth* to *temperature*). The case of human decision making is no different. What is perhaps distinct is the co-existence of multiple possible definitions. Take the case of ‘risk’, which can be defined as tracking the probability of the worst-possible outcome, or the variance among possible outcomes with known probabilities, to name just two possible definitions (see Levy, 2015, Chap. 1; see also Konovalova and Pachur, 2021). The same issue applies to ‘extinction events’, in part due to the

wide range of domains to which the term could be applied (e.g., health, finance, climate). In the present work, we consider cases of repeated choice. Participants are required to repeat the same decision (between an option containing the possibility of extinction and one without this possibility) a maximum of 100 times. Within these scenarios, there are two possible types of extinction events. The first we define as *Keep*, which establishes extinction as an end to the possibility of gains while keeping any gains made thus far. As a real-world example, consider a boxer who risks career-ending injuries every time they take on a high quality opponent for a high value payday. As devastating as career-ending injuries are, they do not undo the benefits gained from such bouts in the past. Our second definition, *Lose*, is much more drastic, referring to a complete wipeout of gains – past, present, and future. As an example, consider the case of a bank whose investment losses are greater than all of the profits generated in their entire history, which results in them going out of business. These definitions of extinction risk also apply to extinction risks involving life, depending on how one views death. While Alice and Bob might agree that death is an end, Alice might consider that this, nevertheless, does not erase all the good experiences accumulated throughout life (*Keep*). Bob, however, might consider that those experiences are also forgotten and therefore erased (*Lose*). Alice’s view is reminiscent of philosophical positions that underscore the experiences we have collected as fundamental to the meaning of life (Eagleton, 2008; Mitsis, 2020; Seneca et al., 2004), whereas Bob’s view would be reflected in positions such as existential nihilism, which emphasize the futility of human endeavors in light of death (Kuhn, 1992; Sartre, 1972).

These two definitions of extinction mirror debates in the literature on the value of a statistical life and mortality-risk valuation (Kniesner & Viscusi, 2019). Some researchers have argued that the value of statistical life should vary with age—an idea sometimes morbidly labeled the ‘senior death discount’—on the grounds that older individuals have fewer remaining life years (Harris, 1985; Krupnick, 2007; M. Lockwood, 1988; A. Williams, 1997). However, empirical work on whether people’s preferences align with this perspective shows inconsistent findings: stated

preference studies show mixed results (a review by Krupnick (2007) found that 14 of 26 studies supported a ‘senior discount’, while the remaining 12 found no effect or even a ‘senior premium’); in contrast, revealed preference studies suggest an inverse U-shaped pattern, where the lives of both younger adults and seniors are valued lower than the lives of middle-aged adults, likely reflecting confounds such as income (Aldy & Viscusi, 2007; Kniesner & Viscusi, 2019). In policy applications, using a senior discount is highly contentious and usually a uniform value of life is applied independent of age (Kniesner & Viscusi, 2019; Krupnick, 2007). Retaining both the Lose definition of extinction (which reflects roughly similar costs of extinction independent of age) and the Keep definition (which reflects reducing costs with increasing age) makes our task applicable to views that involve a senior discount and views that assume age-independent valuations of life.

Both definitions of extinction can be implemented in the context of a repeated risky-choice task. Consider a lottery problem with a ‘risky’ and a ‘safe’ option:

$$\text{Risky} = \begin{pmatrix} r_1 & r_2 & r_E \\ p_1 & p_2 & p_E \end{pmatrix} \quad \text{Safe} = \begin{pmatrix} s_1 & s_2 \\ q_1 & q_2 \end{pmatrix}$$

Each lottery is comprised of mutually exclusive outcomes (e.g., r_1) that occur with a given probability (e.g., p_1). Note that $p_1 + p_2 + p_E = q_1 + q_2 = 1$. The outcome r_E corresponds to the extinction event. The consequences of this event will depend on the definition adopted.

To give a more concrete example:

$$\text{Risky} = \begin{pmatrix} \text{£10} & \text{£0} & \text{Extinct} \\ .475 & .475 & .05 \end{pmatrix} \quad \text{Safe} = \begin{pmatrix} \text{£1} & \text{£0} \\ .50 & .50 \end{pmatrix}$$

If a decision maker only encounters a single lottery problem once, then there is no difference between the potential £0 outcome from the Safe option and the Extinct outcome from the Risky option. For the two outcomes to become distinguishable, it is necessary for the decision maker to engage with multiple lottery problems and

to accumulate gains. In the present implementation of the EGT, they will encounter the same lottery problem multiple times. Knowledge of how many times a lottery problem will be encountered allows the decision maker to better evaluate the exposure to risk vis-à-vis the opportunity for maximizing gains. This evaluation depends on the operating definition of extinction. In the next section, we outline the optimal strategies for both the Keep and Lose definitions of extinction.

Optimal Solution for the Keep Definition

The Keep definition of extinction captures the idea of an end to the accumulation of gains, without any loss of past gains. In the EGT, where a lottery problem is encountered multiple times, this definition implies that the expected payoff of a strategy strongly depends on the position in the sequence of trials where the risky choices are played. Early in the experiment, extinction carries a higher (opportunity) cost than it does later in the experiment, as one misses out on more opportunities to earn money. By contrast, in the final trial, the opportunity cost is zero and the cost of drawing the extinction option is equal to the £0 outcome. Therefore, it would make sense to play the safe lottery in the first trials and the risky lottery towards the end.

Specifically, we show that the expected value is maximised by following a strategy with a single switch point, before which participants should choose the safe option and after which they should choose the risky option. The expected value (\mathbb{E}) for the Keep condition, assuming that a participant follows this optimal order of choices (that is, they first make N_{safe} safe choices and afterwards N_{risky}

risky choices for a fixed total of $N_{\text{safe}} + N_{\text{risky}} = N_{\text{total}}$ choices), is given by

$$\begin{aligned}
\mathbb{E}(N_{\text{risky}}) = & \underbrace{\bar{s} \times N_{\text{safe}}}_{\text{Expected value of safe-choice trials}} \\
& + p_E \times \underbrace{\sum_{i=0}^{N_{\text{risky}}-1} (1 - p_E)^i \times i \times \bar{r}}_{\text{Expected value of the risky-choice trials if one goes extinct at some point}} \\
& + \underbrace{(1 - p_E)^{N_{\text{risky}}} \times N_{\text{risky}} \times \bar{r}}_{\text{Expected value of the risky-choice trials if one does not go extinct}},
\end{aligned} \tag{5.1}$$

where \bar{s} denotes the expected value of choosing the safe lottery, and \bar{r} denotes the expected value of choosing the risky lottery – extinction excluded – and p_E denotes the probability of going extinct when playing the risky lottery. If the safe choices are made first, then the expected value from the safe-choice trials is the number of safe-choice trials multiplied by the expected value per trial (first line of Equation 5.1). The second line denotes the probability of going extinct in the risky-choice trial multiplied by the payoff of this type of extinction (i.e., the expected value from the risky-choice trials over all extinction outcomes). Finally, the third line denotes the probability of surviving the entire experiment multiplied by the payoff of the risky-choices in this case (in the main text we derive all optimal solutions on the basis of expected value; the qualitative patterns of the optimal solutions are very similar if one instead uses expected utility, although with somewhat fewer risky choices, see Appendix C.2).

The optimal number of risky choices can then be obtained by finding the maximum expected value across all possible N_{risky} ,

$$\operatorname{argmax}_{N_{\text{risky}}} \mathbb{E}(N_{\text{risky}}). \tag{5.2}$$

But what if participants do not follow the optimal ordering? If the risky-choice trials are not played strictly after the initial block of N_{safe} safe-choice trials, then

the expected value for the latter (first term of Equation 5.1) changes. For example, consider the case where we replace two safe choices from this block with risky choices, namely the j th and k th trials, with $j < k < N_{\text{safe}}$. Then

$$\begin{aligned}
\bar{s} \times N_{\text{safe}} &\xrightarrow{\text{replaced with}} \underbrace{\bar{s} \times (j-1)}_{\text{first } j-1 \text{ safe choices}} \\
&+ \underbrace{(\bar{s} \times (k-j) + \bar{r}) \times (1-p_E)}_{\text{risky choice at } j \text{ and safe choices between } j+1 \text{ and } k-1} \\
&+ \underbrace{(\bar{s} \times (N_{\text{safe}} - k - 1) + \bar{r}) \times (1-p_E)^2}_{\substack{\text{risky choice at } k \\ \text{and safe choices between } k+1 \text{ and } N_{\text{safe}}}}.
\end{aligned} \tag{5.3}$$

Comparing Equation 5.1 and Equation 5.3, we can see why the expected value is reduced whenever one deviates from the optimal order of playing all safe lotteries first and then all risky lotteries. The expected value for some of the safe choices is multiplied by the probability of survival (a number smaller than one, here $(1-p_E)$ and $(1-p_E)^2$).

Finally, consider the case that the decision maker is given some initial endowment. Then, the expected value for both choice options would simply change by a constant, namely the value of that endowment. This illustrates a more general fact: the optimal strategy for the Keep definition does not depend on initial earnings, as the decision maker can keep them regardless of their choice in the next trial.

Optimal Solution for the Lose Definition

While it is relatively obvious that one should play risky in the final trial under the Keep definition, it is not so clear in the Lose definition, where one might lose all earnings gained on all previous trials. Indeed, under the Lose definition, the optimal solution is *dynamic* and has to account for participants' luck (i.e., their winnings from previous trials). For example, if a participant chooses the risky option in the first three trials, they may receive the maximum payoff three times, or they could receive zero payoff three times. In the first case, they have more to lose in subsequent risky choices than in the second case, and therefore they should play less risky in

the following trials.

Deriving the dynamic optimal solution for this scenario requires the application of the *Bellman equation*, a standard method in economics (e.g., Dixit, 1990). In particular, this method uses backward induction by using the relationship between the value function at one trial and the value function in the next trial.

Let \mathcal{Z} denote current earnings. We can estimate the expected value for the total game following a risky choice with N remaining trials as

$$\begin{aligned}\mathbb{E}_{\text{risky}}(N, \mathcal{Z}) &= p_1 \times \mathbb{E}(N - 1, \mathcal{Z} + r_1) \\ &+ p_2 \times \mathbb{E}(N - 1, \mathcal{Z} + r_2),\end{aligned}\tag{5.4}$$

with

$$\mathbb{E}_{\text{risky}}(0, \mathcal{Z}) = \mathcal{Z}.\tag{5.5}$$

Analogously, we can estimate the expected value of the total game following a safe choice,

$$\begin{aligned}\mathbb{E}_{\text{safe}}(N, \mathcal{Z}) &= q_1 \times \mathbb{E}(N - 1, \mathcal{Z} + s_1) \\ &+ q_2 \times \mathbb{E}(N - 1, \mathcal{Z} + s_2).\end{aligned}\tag{5.6}$$

The optimal choice in a given trial is then risky if $\mathbb{E}_{\text{risky}}(N, \mathcal{Z}) > \mathbb{E}_{\text{safe}}(N, \mathcal{Z})$ and otherwise it is safe. Finally, the expected value of the whole game, when following the optimal dynamic strategy, is determined by the maximum of $\mathbb{E}_{\text{risky}}(N, \mathcal{Z})$ and $\mathbb{E}_{\text{safe}}(N, \mathcal{Z})$, given that an optimal agent would simply play whichever gamble has higher expected value,

$$\mathbb{E}(N, \mathcal{Z}) = \max(\mathbb{E}_{\text{risky}}(N, \mathcal{Z}), \mathbb{E}_{\text{safe}}(N, \mathcal{Z})).\tag{5.7}$$

Unlike the Keep variant, the Lose variant does not imply a single optimal switch point. Instead, there are different possible strategies that would approximate the optimal solution well. Following the optimal policy closely would imply a gradual decrease of the number of risky choices; however, strategies that start safe

or risky and then switch to risky or safe, or a strategy playing mostly safe throughout the experiment would lead to similar expected value, as long as the total number of risky choices is similar. The reason being that, ultimately, the most important factor that determines the payoff in the Lose variant is the total number of risky choices played and, due to the large number of trials, this number is affected relatively little by luck. This is also reflected by the fact that a non-dynamic optimal strategy (i.e., a version of the optimal strategy where all choices are specified in advance and which does not allow taking the current endowment into account) leads to payoffs that are relatively close to payoffs from the dynamic solution (see Appendix C.3).

5.1.2 Experiment 1: Empirical Choice Patterns for the Lose and Keep Conditions With 100 Trials

In the previous sections, we 1) argued that the extant literature makes conflicting predictions regarding whether participants would be risk-seeking or risk-averse when deciding whether to engage in actions that risk extinction, and 2) derived optimal solutions for the EGT. We now report an implementation of the EGT that operationalises the Lose and Keep definitions of extinction and compare participants' choices against the optimal strategy.

Method

Participants & Power. Participants were paid £1.50 for a 10-minute study. On top of that base payment, participants in the Lose condition earned an average bonus of £0.50, interquartile range (IQR) [£0, £0.77], and in the Keep condition £1.01, IQR [£0.39, £1.39]. The study was approved by the ethics chair of UCL's Department of Experimental Psychology (EP/2021/005). All experiments in this section were hosted online on a JATOS (Lange et al., 2015) server, and participants for all experiments were recruited via Prolific, which has been shown to have high data quality compared to other crowdsourcing vendors (Douglas et al., 2023; P. Eyal et al., 2021; Stagnaro et al., 2024). Initially, 196 participants signed up for the study. We excluded participants who failed one or more of four different comprehension and

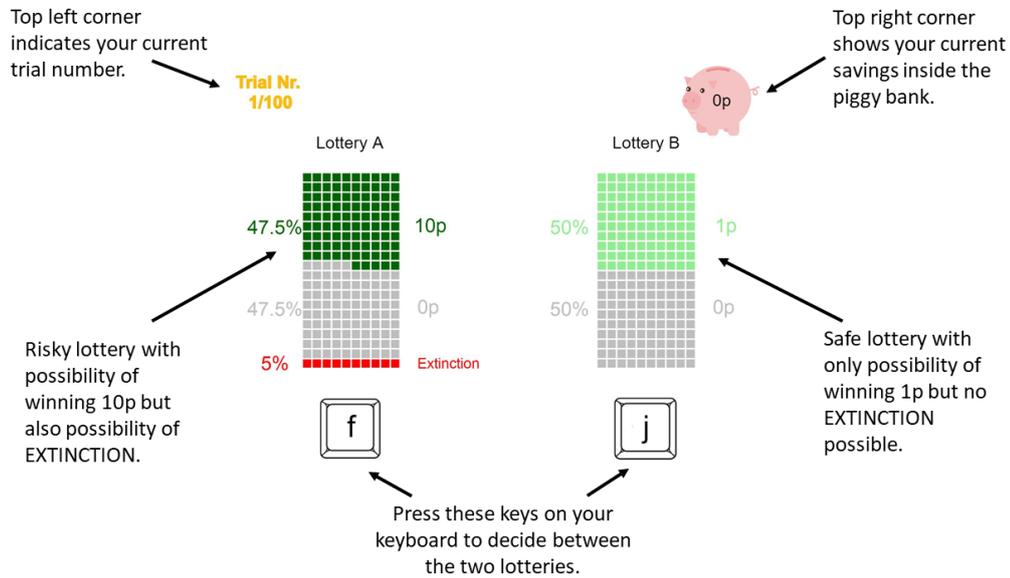
attention checks: two participants indicated the wrong number of trials, four participants did not indicate correctly that the piggy bank in the top right corner of the screen showed their current bonus earnings, four participants indicated that they did not remember the possible payoffs in the experiment, and 33 participants indicated that they did not understand the nature of the extinction event. Some participants failed multiple of these checks. After these exclusions, we obtained a final sample of 157 participants (55 female, 102 male; mean age = 38). 90 participants were in the Keep condition and 67 participants in the Lose condition. A simulation-based power analysis showed close to 100% power with this design to detect a condition difference of 1 on the logit scale, equivalent to a $\approx 17\%$ difference¹ in the proportion of risky choices between conditions (see Appendix C.4 for details on the power analysis, which further demonstrates the robustness of our analytical approach to the selection effects introduced by the extinction event).

Design and Materials. The experiment was programmed in lab.js (Henninger et al., 2021) and exported as a JATOS file for hosting. The full experiment is shared on the OSF (<https://osf.io/qhkw6/>). Participants played a sequence of 20 practice trials, followed by 100 incentivised trials. Throughout the experiment, participants could see how many trials they had left and how much money they had currently accumulated. On each trial, participants had to choose between two lotteries: a safe lottery with a 50% chance of not winning anything, and a 50% chance of winning 1p; and a risky lottery with a 47.5% chance of not winning anything, a 47.5% chance of winning 10p, and a 5% chance of extinction (see Figure 5.1 for an example trial). The task parameters were chosen by simulating from the optimal strategy and selecting values where optimality implies playing risky a few times but less than half of the time in both conditions (Lose $\approx 10\%$, Keep 44%). These parameters reflect that, although selecting the risky option can occasionally be advantageous, regularly taking extinction-level risks—especially in the Lose condition—is generally unfavourable (the values can easily be adjusted in future studies, and we modify the payoffs in Experiment 3). In Experiment 1, the risky lottery was always displayed

¹In estimated average riskiness, the main quantity of interest introduced in more detail below. For comparison, the difference in estimated average riskiness was 35% in Experiment 1.

on the left, and the safe lottery on the right (Experiment 3b replicated the findings of Experiment 1, where lottery position was counterbalanced between participants).

Figure 5.1: The Experimental Setup as Shown to Participants on the Instruction Screen



Our two experimental conditions represented the two extinction definitions outlined in the Introduction. In the Keep condition, participants could keep what they had earned if they experienced the extinction event but could not earn additional money in the following trials. In the Lose condition, participants' entire earnings would be wiped out when experiencing the extinction event and they also could not earn anything in future trials. After encountering an extinction event, participants still needed to play the remaining trials (to avoid any incentives for early extinction); however, the piggy bank in the top right corner turned into a mushroom cloud to indicate extinction (we did not analyse choices made after the extinction event was experienced). At the end of the experiment, participants received their earnings as a bonus payment.

Procedure. Participants started the experiment on a welcome page with a high-level description of the experiment and the following two warnings: 'We advise against participation if you have or had problems with gambling (i.e., gambling addiction) at any point in your life,' and 'At one point during the experiment, you may view

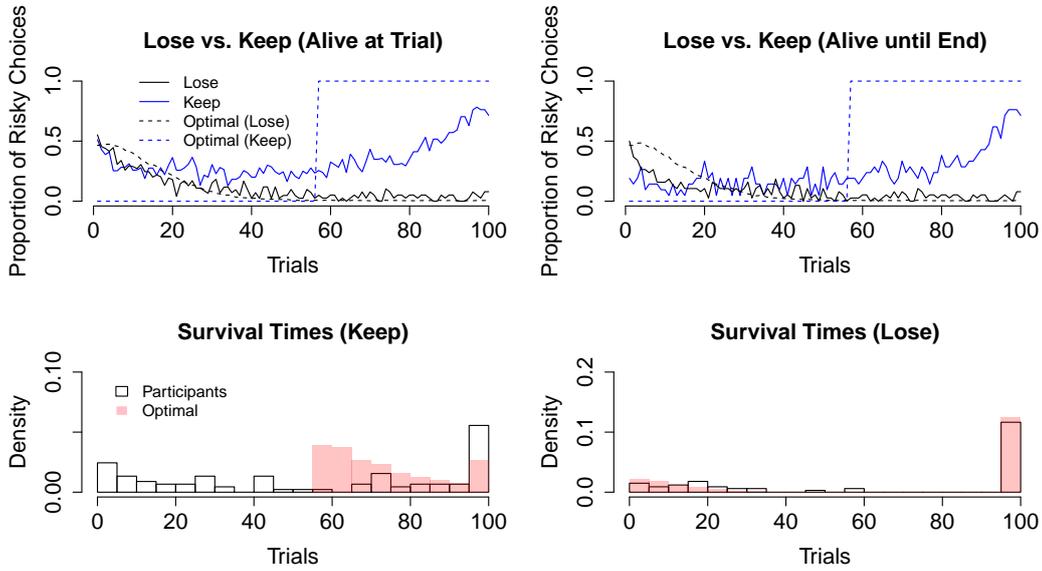
a picture of a mushroom cloud. If you are not comfortable with this, please do not continue.’ After agreeing to the consent forms, participants received detailed task instructions. These instructions included the probabilities and outcomes of all the gambles, what would happen if they drew the extinction event, and an annotated illustration of the lottery screen (see Figure 5.1).

After reading all instructions, participants made decisions in 20 practice trials. Some participants drew an extinction event during these practice trials. In this and the following experiments only a minority of participants experienced the extinction event during practice (Experiment 1: 43 out of 157, Experiment 2: 63 out of 173, Experiment 3a 46 out of 147 participants, Experiment 3b: 116 out of 305 participants), and we found no difference in the proportion of risky choices based on whether or not participants drew the extinction event during the practice trials (Experiment 1: $\chi^2(1) = 0.31, p = .575$; Experiment 2: $\chi^2(1) = 0.78, p = .376$; Experiment 3a: $\chi^2(1) = 0.96, p = .328$; Experiment 3b: $\chi^2(1) = 0.84, p = .361$).

After the practice trials, participants answered two attention checks: ‘What happens if you choose the risky lottery and you draw the extinction option?’ (correct answer for the Keep condition: ‘I will keep all my past bonus money, but I cannot make more bonus money in the next trials,’ correct answer for the Lose condition: ‘I will lose all my bonus money, and I cannot get any more bonus payments for future trials.’), ‘What does the piggy bank on the top right indicate?’ (correct answer: ‘Your current bonus’). Participants continued to the main part of the experiment independent of their answers to these questions; however, those who failed the attention checks were excluded from data analysis.

In the main task, participants chose between the lotteries 100 times. After all trials were completed, we asked participants two further attention checks before they could finish the experiment: ‘How many trials were there in the main part of the experiment?’ (Correct answer: ‘100’), ‘What were the possible bonus payments per trial in this study?’ (Correct answer: ‘Extinction, 0, +1, +10’; the other incorrect choice options also included ‘extinction’ to avoid any uncertainties about whether ‘extinction’ counts as a bonus payment or not).

Figure 5.2: Risky Choices and Survival Times in Experiment 1



Note. The optimal solution for the Lose condition is based on simulating 5000 participants who follow the dynamic solution (Equations 5.4–5.7). The decision is executed based on a softmax decision function, as otherwise tiny differences in expected value in the first choices would lead to deterministic switching between the safe and risky options. We use a very low temperature (0.02) so that the solution is very similar to a deterministic decision rule. The same qualitative pattern arises for a range of temperature values. For a version of this figure that shows model predictions from the fitted mixed effects model along with 95% confidence bands see <https://osf.io/af8sr>. See also Figure C.2 for a visualisation of the optimal choices in the Lose condition as a function of trial number and current endowment.

Results

Results of Experiment 1 are shown in Figure 5.2. The optimal solution is represented in the top row of Figure 5.2 as a dashed line. In the Keep condition (blue), the optimal strategy is to switch from choosing the safe lottery to choosing the risky lottery after trial 56. In the Lose condition (black), the aggregate optimal strategy is to start with more risky choices and then reduce the proportion of risky choices throughout. The top row of Figure 5.2 suggests that participants' choices (solid lines) differed between the two extinction conditions in a manner qualitatively consistent with the difference in the optimal solutions. Specifically, risky choices increased across trials in the Keep condition, but decreased across trials in the Lose condition. The top row of Figure 5.2 also suggests that choices were quantitatively closer to the optimal solution in the Lose condition than in the Keep condition.

The bottom row of Figure 5.2 shows the survival times of participants (white bars, where 100 trials indicates that the participant survived until the end of the experiment), as well as the survival times expected for participants following the optimal strategy (red bars). This shows that more participants went extinct in the Keep compared to the Lose condition. Again, this indicates sensitivity to the core task characteristics, as extinction is more costly in the Lose condition. Further, in the Keep condition, some participants went extinct earlier than implied by the optimal strategy, whereas in the Lose condition, the survival times tracked the optimal expectations remarkably well (see also Figure C.2 for a visualization of the optimal choice for the Lose condition as a function of trial number and money, and Figure C.6 for a by-participant summary of the data).

To statistically test people's sensitivity to the different extinction conditions, we applied a logistic mixed effects model which predicts choosing the risky option as a function of trial number (with both a linear and a quadratic effect, to account for a potential U-shape), and condition (Keep vs. Lose), as well as an interaction between trial number and condition. The optimal strategy implies a higher proportion of risky choices overall in the Keep condition. We perform this comparison across conditions using the *estimated average riskiness* (i.e., the predicted riskiness based on the model's fixed and random effects assuming each participant responded to all trials, which is then averaged per condition; this measure takes potential non-linearities into account and addresses the selective dropout of more risky participants). For more details on our approach, including a simulation study to verify its accuracy, see Appendix C.4 as well as the supplementary materials. We found a significant difference in the estimated average riskiness between conditions, $z = 6.01$, $p < .001$, with the estimated proportion of risky choices being 52%, 95% CI [43%, 62%] in the Keep condition, as opposed to 17%, 95% CI [10%, 24%] in the Lose condition. For comparison, if all participants had followed the optimal strategy, these proportions would be 44% and 10%.

Further, the optimal strategy implies increasing risky choices during the task for the Keep condition and decreasing risky choices for the Lose condition. In line

with this, we found a significant interaction between trial number (linear effect) and condition, $\chi^2(1) = 54.66, p < .001$, with a positive marginal linear slope in the Keep condition $b = 3.68, 95\% \text{ CI } [1.29, 6.08]$, and a negative marginal linear slope in the Lose condition $b = -6.41, 95\% \text{ CI } [-8.02, -4.79]$. Overall, these results suggest that participants were sensitive to the differences between conditions in a way that was (qualitatively) consistent with the optimal strategy.²

Computational Modelling of Individual Level Strategies

The aggregate pattern of results in the previous section likely arose from a mixture of different strategies on the level of individual participants. For example, it is plausible that the steady increase in the probability of choosing risky in the Keep condition (top-row of Figure 5.2) is actually the result of different participants switching from choosing safe to choosing risky at different time points. Specifically, in the EGT participants can: 1) play the risky option with a constant probability throughout; 2) gradually increase or decrease the probability of playing the safe or risky option; or 3) switch between playing safe and playing risky. These strategies could be played in either a deterministic or probabilistic way (e.g., switching from always safe to always risky vs. switching from playing safe 90% of the time to playing risky 90% of the time).

Modelling the individual-level strategies confers several benefits. First, some of these strategies are qualitatively closer to the optimal solution than others. For example, in the Keep condition, the optimal solution would imply a sudden switch from safe to risky (rather than a constant increase or a reduction), whereas switching from risky to safe, or playing only risky, would result in a lower expected payoff. Therefore, modelling individual-level strategies helps us understand to what extent participants recognized the optimal strategy and were only impeded by their ability to estimate the correct parameters (e.g., some participants may have recognized that a single switch strategy was optimal, but did not know where the optimal switch

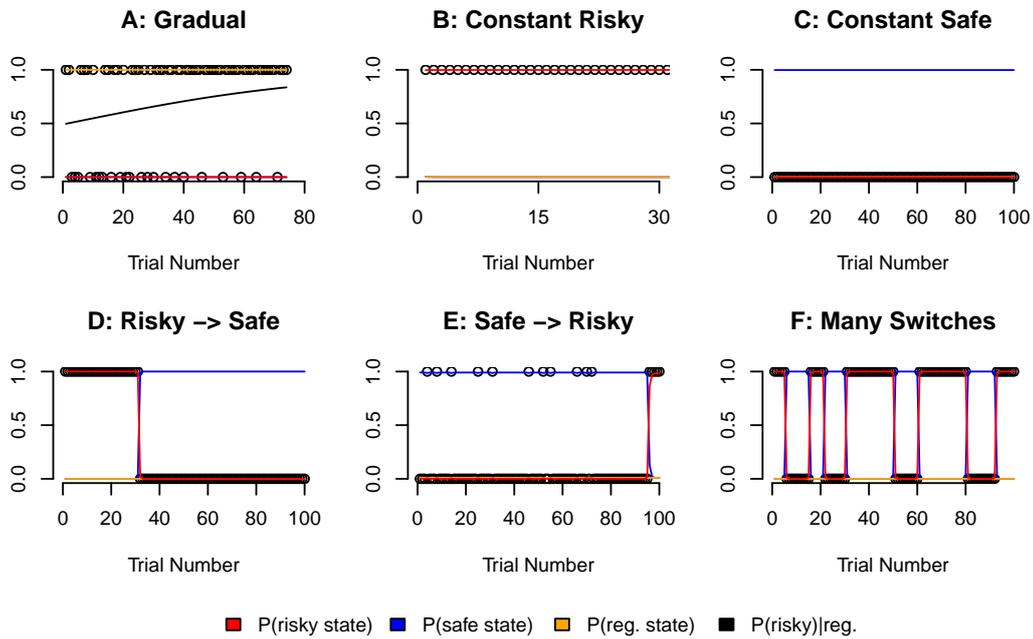
²There was also a significant interaction between trial number (quadratic effect) and condition, $\chi^2(1) = 10.48, p = 0.001$ (Keep: $b = 11.64, 95\% \text{ CI } [4.70, 18.59]$; Lose: $b = -0.365, 95\% \text{ CI } [-5.70, 4.97]$).

point was), or did not even qualitatively recognize the optimal strategy. Second, the optimal strategies differ between conditions in systematic ways. For example, in the Keep condition, it is best to switch from playing safe to playing risky, whereas this strategy is less strongly implied in the Lose condition. Therefore, modelling individual level strategies allows us to better understand differences between conditions and sensitivity to the type of extinction event. Third, we can use parameters from computational models to get a better understanding of specific elements of participants' behaviour. For example, comparing the model estimated switch point to the optimal switch point gives some evidence of whether participants are risk-averse or risk-seeking. In the next sections, we first outline the model specification and then illustrate the three benefits discussed here through application of the model to the experimental data.

Model Specification

To describe the above strategies, we implemented a hierarchical (dependent) mixture model with three components: a *safe state*, in which participants would play mostly the safe choice; a *risky state*, in which participants would play mostly the risky choice; and a *gradual state*, which describes a gradual increase or decrease of the probability of playing risky throughout the experiment (implemented with logistic regression). The model allows participants to switch between the safe and risky states during the experiment to accommodate the different types of strategies outlined above (e.g., switching from safe to risky, or switching from risky to safe). However, participants can not switch away from or into the gradual state. Further, we constrained the states so that the probability of playing risky when in the risky state is always > 0.8 , and the probability of playing risky when in the safe state is always < 0.2 (for a detailed model specification, including information on prior distributions, see Appendix C.5).

Figure 5.3: Six Different Strategies Identified by the (Dependent) Mixture Model



Note. Each panel illustrates one participant who is representative of one of the six strategies. The y-axis denotes the probability of a given state shown in a coloured line (red, blue, and orange lines), the probability of choosing risky given the regression state (black line), and the actual choices as circles (1 denoting a risky choice and 0 denoting a safe choice). When one strategy has a posterior probability of almost 100%, the lines for the other two strategies will overlay at the bottom of the plot (i.e., only one is visible).

Computational Modelling Results

The model showed good convergence (all $\hat{R} < 1.01$, Vehtari et al., 2021) and recovered the patterns in the data well (see Figures C.4 and C.5 for posterior predictives). The computational modelling allowed us to identify six types of individual strategies, as visualised in Figure 5.3:

1. **Gradual:** Participants who continuously increased or decreased their probability of playing risky during the experiment.
2. **Constant Risky:** Participants who mostly played the risky option, and their probability of choosing risky did not change during the experiment.
3. **Constant Safe:** Participants who mostly played the safe option, and their

probability of choosing safe did not change during the experiment.

4. **Risky → Safe:** Participants who switched from playing mostly risky to playing mostly safe at a certain trial number.
5. **Safe → Risky:** Participants who switched from playing safe to playing mostly risky at a certain trial number.
6. **Many Switches:** Participants who switched between playing mostly safe and playing mostly risky multiple times during the experiment.

Figure 5.4 shows that the proportion of strategies employed by different participants varied between conditions, $\chi^2 = 18.24, p = .002$.³ This again indicates sensitivity to the core task characteristics and qualitative understanding of the optimal strategies in several ways. First, participants in the Keep condition switched from safe to risky more often than in the Lose condition, as implied by the optimal strategy. Second, participants in the Lose condition played the safe strategy considerably more often than participants in the Keep condition. This is in line with the optimal strategy: extinction is more costly in the Lose condition than in the Keep condition; consequently, in the Lose condition, playing fewer risky choices is optimal.

In both conditions, the majority of participants were assigned to the gradual strategy (Figure 5.4), which was implemented with logistic regression. Based on the optimal strategies, we would expect the regression to have a positive slope in the Keep condition and a negative slope in the Lose condition.⁴ In line with this, all regression slopes were negative in the Lose condition, and positive in the Keep condition (Figure 5.5, Panel A).

Finally, for the Keep condition, we compared the switch point estimated by the model to the optimal switch point to assess whether those participants who under-

³We use a simulation-based chi-square test as the chi-square approximation may be incorrect given the number of observations per cell. Therefore, no degrees of freedom are provided.

⁴Although the optimal strategy predicts a negative slope for the Lose condition, this prediction is not as strict as the positive slope predicted in the Keep condition. As discussed above, the expected value may be very similar for other trajectories in the Lose condition as long as the total number of risky choices is similar to the one implied by the optimal strategy.

Figure 5.4: Proportion of Participants Allocated to Each of the Six Strategies in Experiment 1

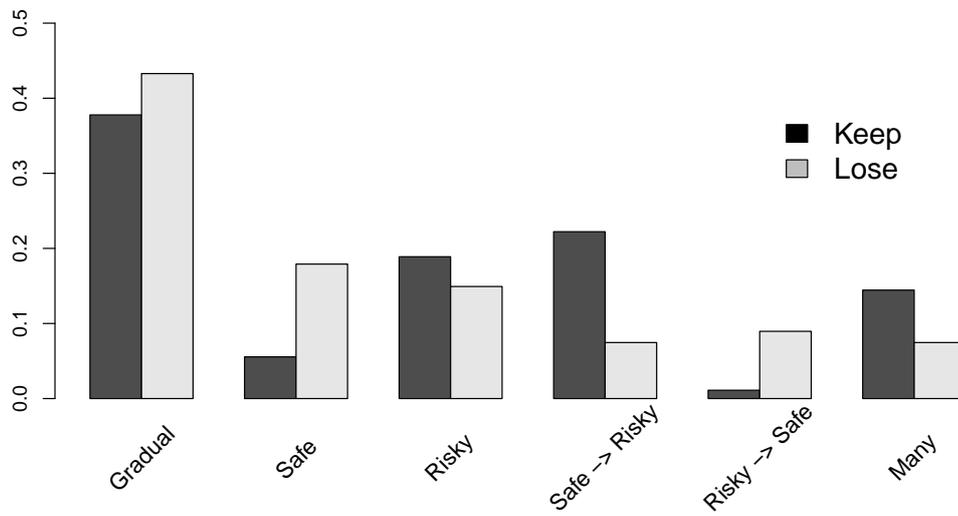
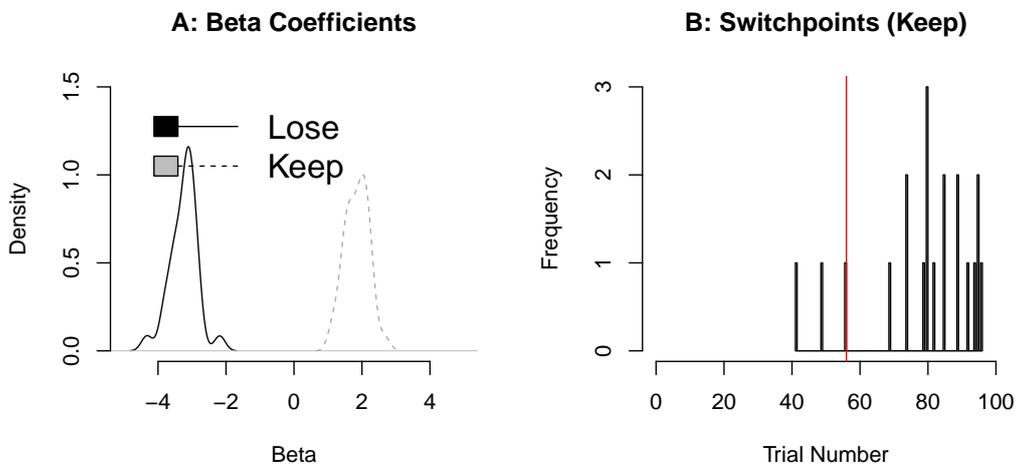


Figure 5.5: Parameter Estimates from the Computational Models in Experiment 1



Note. Panel A shows beta coefficients for both conditions and Panel B shows switch points for the Keep condition. Each panel shows the distribution of posterior mean estimates across participants. Beta coefficients are only plotted for those participants who were assigned to the gradual strategy and switch points are only shown for those participants who were assigned to the Safe → Risky strategy.

stood that the optimal strategy involved a single switch from safe to risky tended to switch too late or too early (Figure 5.5, Panel B). This analysis indicated that most people switched later than the optimal switch point, $t(19) = 6.74, p < .001$,

consistent with risk-aversion.

Discussion

Experiment 1 compared participants' behaviour to the optimal strategies in the Keep and Lose conditions of the EGT. We found that participants' aggregate behaviour was qualitatively in line with the optimal strategies, approximating the optimal solution in the Lose condition particularly well. Computational modelling of individual differences revealed a variety of different strategies employed by individual participants. Even though only a small proportion of participants followed the optimal strategy exactly (i.e., single switch or gradual), most participants used a strategy that at least qualitatively resembled the optimal approach (e.g., constant increase of $p(\text{risky})$ in the Keep condition).

5.1.3 Experiment 2: Introducing a Reset Condition

In Experiment 2, we introduce a new experimental condition – *Reset*. In this condition, participants lose all their earnings when the extinction event occurs, but they can continue playing and accumulate new gains (and potentially draw the extinction event again). In practical terms, Reset is a mirror version of the Keep condition implemented in Experiment 1 (in which participants can keep all the money they have earned, but cannot make new gains after extinction).

The motivation for this new condition is twofold. First, it can be mapped onto real-world risk scenarios, where one loses everything but can recover (e.g., if one's house gets destroyed in a natural disaster). Second, it allows us to test the influence of opportunity cost neglect. Opportunity costs are foregone benefits that one would have incurred if one had chosen an alternative option. A key feature of extinction events is that they have high opportunity cost: all the benefits one could have incurred after the extinction event are now lost. Research shows that people tend to exhibit opportunity cost neglect; that is, they insufficiently take into account opportunity costs as compared to 'direct' costs (Frederick et al., 2009; Persson & Tinghög, 2020; Spiller, 2019).

Comparing the choice patterns in the Keep and Reset conditions of the EGT allows us to make inferences about how well participants can reason about opportunity costs versus ‘direct’ costs. In the Keep condition, a participant’s endowment is not at stake, but their opportunity to keep playing in later trials is at stake (i.e., extinction leads to opportunity cost). Conversely, in the Reset condition, a participant’s endowment is at stake, but their opportunity to continue playing in later trials is not at stake (i.e., extinction leads to ‘direct’ costs). In particular, we can compare the choices at the beginning of the task in the Keep condition (where extinction would lead to high opportunity cost but low direct cost) versus at the end of the task in the Reset condition (where extinction would lead to a high ‘direct’ cost but low opportunity cost). If participants display opportunity cost neglect, we would expect them to play too risky at the beginning of the task in the Keep condition, but we would *not* expect them to play too risky at the end of the task in the Reset condition. Consequently, we might expect more people in the Reset condition to follow the optimal single switch strategy than in the Keep condition (where some people would start too risky).

Overall, Experiment 2 has one main hypothesis and one research question.

1. **Sensitivity to Extinction Condition:** In the Keep condition, participants will increase the proportion of risky choices as they proceed through the experiment, whereas in the Reset condition, they will reduce the proportion of risky choices (preregistered).⁵
2. **Opportunity Cost:** Are participants in the Reset condition more in line with the optimal policy than participants in the Keep condition? (not preregistered)⁶

⁵The preregistration is available at <https://osf.io/8uas6>. Note that in the preregistration, we use a previous terminology and refer to the Keep condition as the Keep-stop condition and the Reset condition as the Lose-continue condition.

⁶We initially preregistered that the main test for the role of opportunity cost neglect is whether there is a significant difference in the number of risky choices between both conditions. However, as this test was based on a static optimal policy for the Reset condition, this is no longer the most informative test (see section ‘Optimal Strategy for the Reset Condition’). Following the preregistration with a test comparing the average riskiness between conditions, would come to the same conclusions.

Optimal Strategy for the Reset Condition

Because participants can lose all earnings accumulated so far in the Reset condition, the optimal strategy depends on how lucky participants have been and how many earnings they have won (as in the Lose condition). The optimal strategy for the Reset condition therefore also follows a dynamic solution. It is the same as in the Lose condition, with the addition that the optimal solution now also factors in the possibility of ‘resets’ (i.e., an event where one loses one’s endowment but can continue playing). Specifically, the expected value of a risky choice with N remaining trials corresponds to

$$\begin{aligned}\mathbb{E}_{\text{risky}}(N, \mathcal{Z}) &= p_1 \times \mathbb{E}(N - 1, \mathcal{Z} + r_1) \\ &+ p_2 \times \mathbb{E}(N - 1, \mathcal{Z} + r_2) \\ &+ p_3 \times \mathbb{E}(N - 1, 0).\end{aligned}\tag{5.8}$$

Apart from this modification, the optimal solution follows the Lose condition—Equation 5.6 determines the expected value of a safe choice, the value when zero trials are left equals the endowment (Equation 5.5), and the choice between safe and risky is determined by whichever of the two options has higher expected value (Equation 5.7).

Even though the Reset condition can conceptually be viewed as a mirror version of the Keep condition (as we argue in the introduction of this experiment), the optimal solution in the Reset condition implies a larger proportion of risky choices than in the Keep condition for the same maximum trial number, probabilities, and outcomes (67.6% vs. 44%). The reason for this is that risky trials in the Reset condition can be played both before and after an extinction event is experienced. Consider a participant in the Reset condition who starts risky and is reset to zero after 40 trials. This participant should then continue playing risky until the endowment becomes so large that it has a higher expected value to play safe rather than risky, which is for much longer than 4 more trials (44 trials is the optimum in the

Keep condition).⁷

Participants

Participants were paid £1.20 for the 8-minute study. On top of that base payment, participants in the Reset condition earned an average bonus of £1.22 IQR [£0.79, £1.57] and in the Keep condition £1.03 IQR [£0.40, £1.35]. The study was approved by the ethics chair for UCL's Department of Experimental Psychology (EP/2021/005). Initially, 197 participants completed the study. One participant indicated the wrong number of trials, six participants did not indicate correctly that the piggy bank in the top right corner of the screen showed their current bonus earnings, six participants did not remember the possible payoffs in the experiment correctly, and 20 participants indicated that they did not understand the nature of the extinction event (some participants failed multiple of these checks). After these exclusions, we obtained a final sample of 173 participants (77 in the Reset condition and 96 in the Keep condition). The mean age of participants was 40.60 years (SD = 12.15). Eighty-eight participants were female and 85 male.

Materials and Procedure

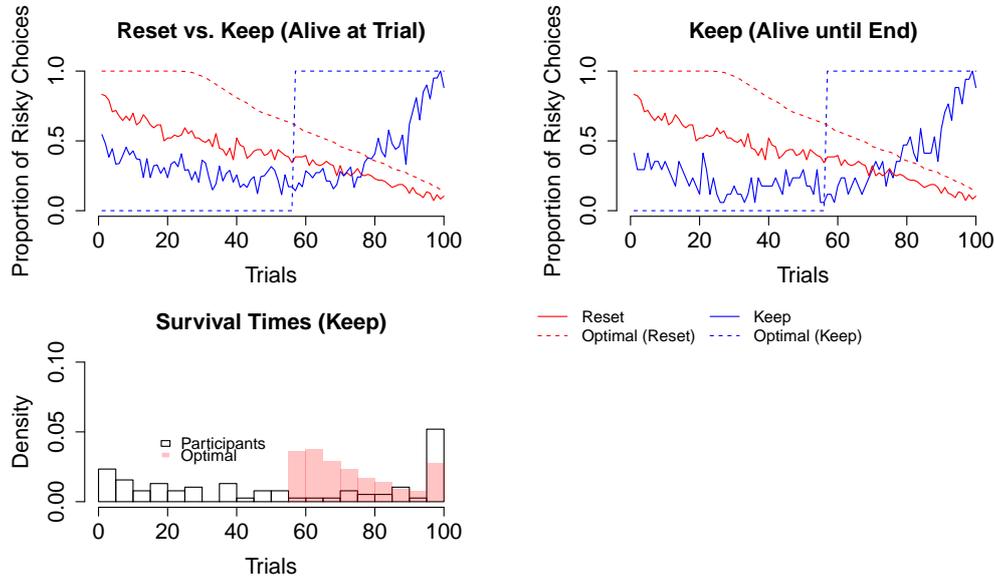
The materials and procedure were the same as in Experiment 1, except that we replaced the Lose condition with the Reset condition. In addition, the position of the risky and safe choices on the screen was counterbalanced between participants.

Results

As in Experiment 1, the differences in participants' choices across conditions were qualitatively in line with the predictions of the optimal solutions (see Figure 5.6). During the experiment, the proportion of risky choices increased in the Keep condition and decreased in the Reset condition.

⁷This difference in optimal strategy is also why the effect of opportunity cost neglect cannot be directly tested by comparing the proportion of risky choices, but instead requires computational modelling of the individual level strategies.

Figure 5.6: Risky Choices and Survival Times in Experiment 2



Note. The optimal solution for the Reset condition is based on simulating 5000 participants, which follow the dynamic solution (Equations 5.5–5.8). The decision is executed based on a softmax decision function, as otherwise tiny differences in expected value in the first choices would lead to deterministic switching between 0 and 1. We use a very low temperature (0.02) so that the solution is very similar to a deterministic decision rule. The same qualitative pattern arises for a range of temperature values. In the Reset condition, participants cannot drop out; therefore, the ‘Alive at Trial’ and ‘Alive until End’ lines are identical in this condition. For a version of this figure that shows model predictions from the fitted mixed effects model along with 95% confidence bands see <https://osf.io/mrvg7>. Also see Figure C.3 for a visualisation of the optimal strategy in the Reset condition as a function of current endowment and trial number.

We again use a mixed effects model with the same fixed effects specification as in Experiment 1 (effects of condition, linear and quadratic effect of trial number, as well as the interactions of condition with trial number).⁸ Consistent with the predictions of the optimal strategies, we found a significant interaction between trial number (linear effect) and condition, $\chi^2(1) = 1372.28, p < .001$, with a pos-

⁸The model also had by-participant random intercepts and initially by-participant random slopes for both the linear and the quadratic effect of trial number. As the maximal model, showed convergence issues (degenerate Hessian with 4 negative eigenvalues) and candidate models with random slopes for only the linear effect of trial number also did not converge without issues, we report results from a reduced model with only by-participant random intercepts. The maximal model showed the same pattern of statistical significant effects as the reduced model, apart from the main effect of condition as indicated in the main text. In addition to the effects reported in the main text, we also found a significant interaction effect between condition and trial number (quadratic effect), $\chi^2(1) = 157.39, p < .001$; Keep: $b = 9.87, 95\% \text{ CI } [8.77, 10.97]$; Reset: $b = -0.72, 95\% \text{ CI } [-1.48, 0.04]$.

itive marginal linear slope in the Keep condition, $b = 2.44$, 95% CI [2.11, 2.76], and a negative marginal linear slope in the Reset condition, $b = -4.46$, 95% CI [-4.68, -4.24]. We further found a just below threshold significant difference between the conditions in estimated average riskiness, $z = 1.98$, $p = .048$ (when using the maximal model, which showed convergence issues, this difference was no longer significant, $z = 1.29$, $p = .197$), with the estimated proportion of risky choices being 40%, 95% CI [35%, 46%] in the Reset condition and 49%, 95% CI [42%, 56%] in the Keep condition.

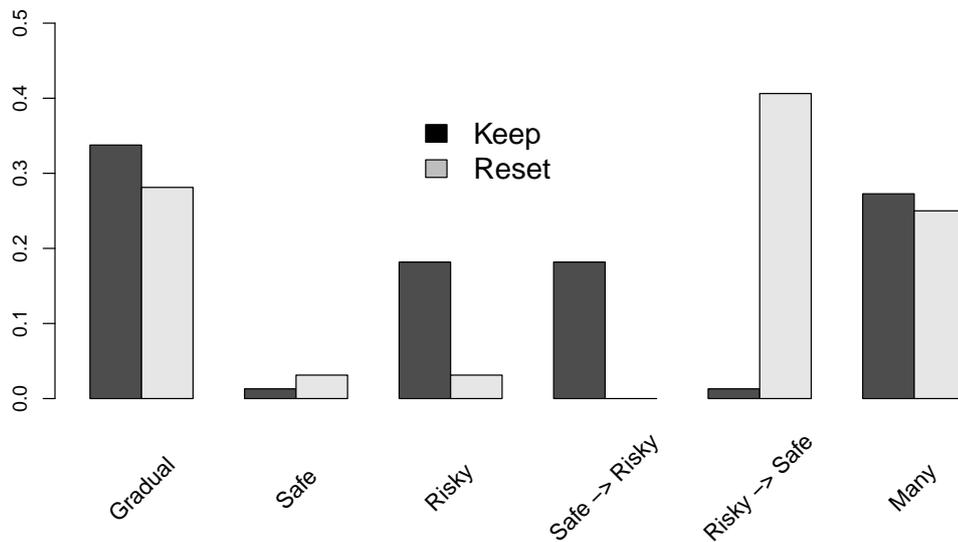
Computational Modelling Results

We applied the same computational model as in Experiment 1 to investigate participants' decision strategies on an individual level. The model again showed good convergence (all $\hat{R} < 1.03$). In the Keep condition, the optimal strategy still involves switching to playing risky after trial 56. In the Reset condition, the optimal strategy involves switching in the other direction (Risky \rightarrow Safe); however, because the solution is dynamic, this switch does not always occur at the same trial. Instead, the timing of the optimal switch point varies between participants. If participants do not take opportunity cost into account, we would expect those in the Keep condition to start more risky than implied by the optimal strategy, and consequently fewer participants in this condition would follow the optimal strategy.

Participants' strategies differed considerably between conditions (see Figure 5.7; $\chi^2 = 57.17$, $p < .001$). In line with the optimal solutions, participants were more likely to switch from risky to safe in the Reset condition, and more likely to switch from safe to risky in the Keep condition. In line with our hypothesis that participants find it easier to reason about 'direct' costs than opportunity costs, more participants followed the optimal strategy (switching from risky to safe) in the Reset condition than in the Keep condition (switching from safe to risky), $\chi^2(1) = 4.80$, $p = .028$.

A comparison of participants' switch points to optimal switch points suggests risk aversion in both conditions, as in Experiment 1. As shown in Figure 5.8, those

Figure 5.7: Proportion of Participants Allocated to Each of the Six Strategies in Experiment 2



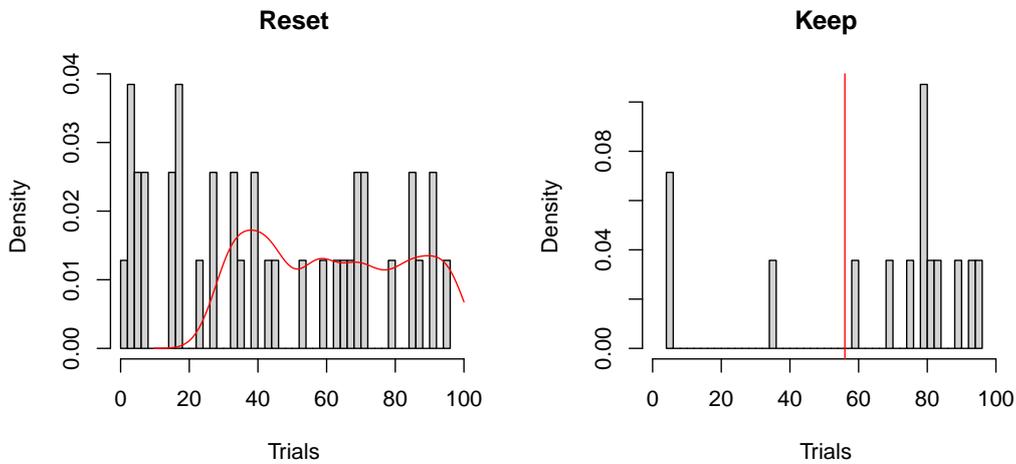
Note. More Participants Switch from Safe to Risky in the Reset Condition than from Risky to Safe in the Keep Condition.

in the Reset condition switched from choosing risky to safe earlier than the optimal solution suggested, $t(38.85) = 3.81, p < .001$, whereas those in the Keep condition descriptively switched from safe to risky later than suggested, though this effect is not statistically significant, $t(13) = 1.32, p = .208$ (this test is based on only 14 participants, three of whom switched very early). As in Experiment 1, the slopes of those participants assigned to the gradual strategy showed the expected direction (mostly negative in the Reset condition and mostly positive in the Keep condition, though there is more variability in the slopes in the reset condition, see Figure C.7).

Discussion

In Experiment 2, we introduced the Reset condition, in which participants lose everything if they draw the extinction event, but can continue to play and acquire future gains. As in Experiment 1, participants demonstrated sensitivity to the core task characteristics and we found some alignment between individual-level strategies and the optimal solution. Furthermore, we also found that more participants recognized the optimal single switch strategy in the Reset than in the Keep condi-

Figure 5.8: Participants' Switch Points and Optimal Switch Points for Experiment 2



Note. Histogram of switch points for those participants that followed a single switch strategy. Red denotes the distribution of optimal switch points in Reset and the single optimal switch point in Keep.

tion. This suggests that it is possibly easier for participants to reason about ‘real’ losses (i.e., their endowment, as represented by money in the piggy bank) than opportunity costs (i.e., not being able to continue playing and earning money).

5.1.4 Experiment 3: What Psychological Factors Shape Decisions Under Extinction Risk?

In the first two experiments, we introduced three different definitions of extinction or catastrophic risks (Keep, Lose, & Reset) and tested the influence of opportunity cost neglect via comparison between the Keep and Reset conditions. Experiment 3 aims to further demonstrate the EGT’s potential by varying features of the task *within* extinction conditions to test the influence of a variety of psychological factors on decision making under extinction risk. In particular, we explore the influence of loss chasing (Gainsbury et al., 2014; Lister et al., 2016), opportunity cost neglect (e.g., Frederick et al., 2009), and scope insensitivity (e.g., Desvousges et al., 1993) on decision making under extinction risk. Our aim is to demonstrate the utility of the EGT for studying the influence of such phenomena. Consequently, the following does not provide a definitive account of these phenomena, but rather

a first demonstration of how they can be investigated using the EGT.

Experiment 3a: Endowment and Losses

In Experiment 3a, we extended the Lose condition by introducing small losses and a starting endowment (Lose + Endowment condition). This new condition is identical to the Lose condition from Experiment 1, with the addition that participants now start with a small endowment of 50p and decide between a safe lottery that has a 50% chance of winning 1p and a 50% chance of *losing* 1p; and a risky lottery that has a 47.5% chance of winning 10p, a 47.5% chance of *losing* 1p, and a 5% chance of extinction. In other words, the 0p outcome from before is now replaced with a $-1p$ outcome.

The Lose + Endowment condition has both practical and theoretical motivations. From a practical perspective, in the real world people usually do not start from zero but with something to lose. In addition, there is usually a possibility of small losses even under the safe option. For instance, someone who decides *not* to jaywalk may risk being late for a meeting. This would likely result in more risky choices because people often play more risky to recover from previous losses (a phenomenon termed ‘loss chasing’ in the literature on casino gambling; Gainsbury et al., 2014; Lister et al., 2016), especially when these losses have not yet been realised (i.e., paper losses or losses that have not yet been ‘cashed out’ in the eyes of the participant; Imas, 2016). Losing money in the piggy bank in our task would likely constitute an unrealised loss as this money has not yet been awarded to the player.

This condition additionally allows us to probe in more detail why participants’ proportion of risky choices declined during the task and, consequently, approximated the optimal solution in the Lose condition remarkably well in Experiment 1 (Figure 5.4). One possible explanation for this decline is that participants play less riskily the higher their endowment (i.e., the more money is displayed in the piggy bank), similar to the idea of opportunity cost neglect investigated in Experiment 2. If this was indeed the case, we would expect fewer risky choices at the start

of the Lose + Endowment condition, where there is a higher starting endowment, than in the normal Lose condition (note that this difference would not be affected by loss chasing, as participants have not yet experienced any losses at this point). The proportion of risky choices should, however, decline more slowly in the Lose + Endowment condition, as participants' endowment will not increase as quickly as in the normal Lose condition due to the $-1p$ outcome.

Therefore, this study has two main research questions:

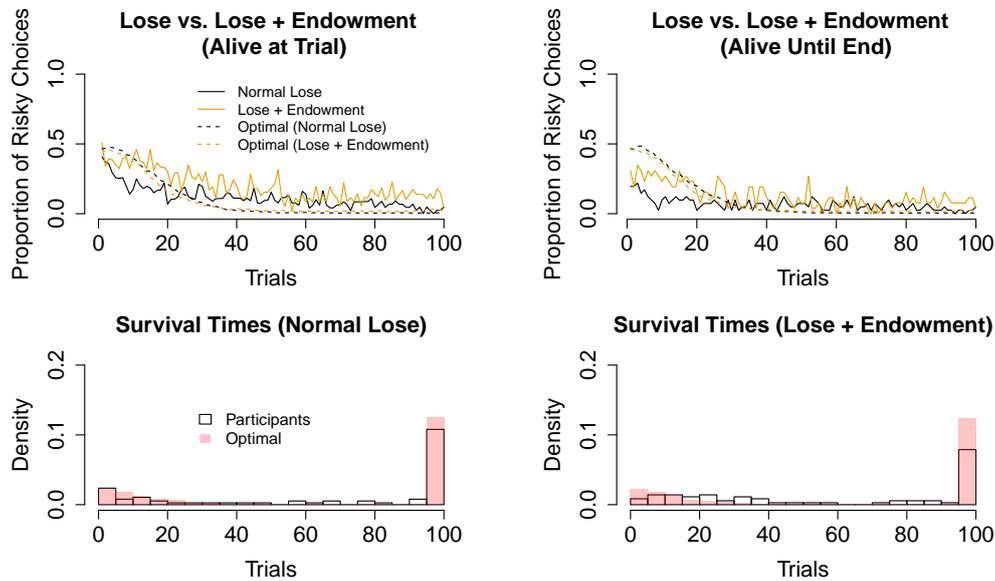
1. Is overall risk-seeking increased in the Lose + Endowment condition compared to the normal Lose condition? (not preregistered).
2. Does the proportion of risky choices decline faster in the normal Lose condition than in the Lose + Endowment condition? (preregistered)⁹

Participants. The study was approved by the ethics chair of UCL's Department of Experimental Psychology (EP/2021/005). Participants were paid £1.20 to participate in an 8-minute study. On top of that base payment, participants in the normal Lose condition earned an average bonus of £0.44 IQR [£0.00, £0.60] and in the Lose + Endowment condition £0.41 IQR [£0.00, £0.63]. Initially, 210 participants signed up for the study. Three participants failed the attention check asking them about the total number of trials, nine failed the check asking about the possible outcomes of the lotteries, 25 did not know what the extinction event indicated, and six did not know what the piggy bank indicated. We also excluded three participants that seemed to have signed up twice, due to a technical problem. After excluding participants who failed these attention checks, we were left with a final sample of 147 (76 in the normal Lose condition and 71 in the Lose + Endowment condition). The average age was 41.21 (SD = 14.73), 72 participants were female and 75 were male.

Materials and Procedure. The normal Lose condition was the same as in Experiment 1. For the Lose + Endowment condition, we updated the instructions and

⁹The preregistration is available at <https://osf.io/8dntk>. If the results of this test were to be statistically significant, we would have followed up by testing the difference at the first trial to delineate this from the influence of loss chasing.

Figure 5.9: Risky Choices and Survival Times for Experiment 3a



Note. The optimal solution is based on simulating 5000 participants who follow the dynamic solution (Equations 5.4–5.7). The decision is executed based on a softmax decision function, as otherwise tiny differences in expected value in the first choices would lead to deterministic switching between 0 and 1. We use a very low temperature (0.02) so that the solution is very similar to a deterministic decision rule. The same qualitative pattern arises for a range of temperature values. For a version of this figure that shows model predictions from the fitted mixed effects model along with 95% confidence bands, see <https://osf.io/g6ck3>.

lottery screens to reflect the possibility of small losses, and the piggy bank on the first page of the experiment started with a value of 50. As in Experiment 2, we counterbalanced the position of the options on the screen.

Results. Figure 5.9 (top row) suggests that the overall proportion of risky choices was higher in the Lose + Endowment condition than in the original Lose condition. Furthermore, as in Experiment 1, the proportion of people who survived until the end was similar to the optimal solution in the normal Lose condition; however, it was lower than the optimal solution in the Lose + Endowment condition (Figure 5.9, bottom row), indicating greater risk-seeking in the Lose + Endowment condition.

As in the previous experiments, we used a logistic mixed effects model with effects of condition, trial number (linear & quadratic effect), and their in-

teraction. This indicated a significant difference in estimated average riskiness, $z = 2.74, p = .006$, with the estimated proportion of risky choices being 17%, 95% CI [11%, 23%] in the normal Lose condition and 28%, 95% CI [18%, 38%] in the Lose + Endowment condition (for comparison, the optimal would be 10% in both conditions). Further, we found no evidence for an interaction between trial number (linear effect) and condition, $\chi^2(1) = 2.20, p = 0.138$, with a negative marginal slope in both the normal Lose condition, $b = -3.30, 95\% \text{ CI } [-4.78, -1.81]$, and the Lose + Endowment condition, $b = -1.88, 95\% \text{ CI } [-3.61, -0.14]$.¹⁰ These results are consistent with RQ1 (more risky choices after the introduction of Losses) but not RQ2 (faster decrease in risky choices in the normal Lose condition compared to the Lose + Endowment condition).

Computational Modelling Results. The model again showed good convergence (all $\hat{R} < 1.01$). Since the optimal strategy for both conditions was virtually the same in this experiment, we did not expect any major differences in terms of participants' assignment to the different strategies. In line with this, we found no evidence for differences in strategies between conditions, $\chi^2 = 3.50, p = .643$ (see Appendix C.8.3 for a visualisation of the computational modelling results for this experiment).

Discussion. Overall, Experiment 3a provides evidence for risk-seeking after (paper) losses, as participants' responses were more risk-seeking in the condition including small (paper) losses (Lose + Endowment) than in the condition without such losses (Lose). Further, we did not find evidence that participants' risky choices are mainly determined by the amount of money in the endowment rather than the expected value of the total game. Instead, risky choices decreased during the experiment at a similar rate for both the Lose + Endowment and normal Lose conditions. This suggests that factors other than the endowment (e.g., exploration in the beginning of the task) cause the reduction in risky choices during the experiment.

¹⁰We also did not find a significant interaction between trial (quadratic effect) and condition ($\chi^2(1) = 0.04, p = .839$) with a positive slope estimate in both conditions (Lose: $b = 1.69, 95\% \text{ CI } [-1.71, 5.09]$, Lose + Endowment: $b = 2.16, 95\% \text{ CI } [-1.18, 5.51]$).

Experiment 3b: Varying the Maximum Number of Trials

In Experiment 3b, we turn to another important feature of our task: unlike most other gambling tasks, the maximum number of trials has considerable impact on the optimal strategies in the EGT. Intuitively, as the maximum number of trials decreases, less can be lost by drawing the extinction outcome, and consequently participants should choose the risky lottery more often. A similar line of reasoning also applies to real-world decisions under extinction risk: when more is at stake in terms of investment or future time lost from going extinct, it becomes more important to avoid extinction risk. However, research on choice bracketing (Read et al., 2000) and scope insensitivity (Kahneman et al., 1999) suggests that people may not be sufficiently sensitive to the value that could be lost in case of extinction.

Choice bracketing refers to whether people, when faced with a sequence or set of choices, are more likely to integrate across them ('broad bracketing') or to consider each of the choices in isolation ('narrow bracketing'). People tend to view individual choices in isolation rather than integrating across the set; that is, they exhibit narrow bracketing (Bland, 2019; Rabin & Weizsäcker, 2009; Read et al., 2000). If narrow bracketing also pertains to the EGT, it would imply insensitivity to the total number of choices. Scope insensitivity describes a similar phenomenon whereby people do not value a good in proportion to its scope or size (Kahneman et al., 1999). It has been shown in a variety of domains, such as contingent valuation judgment (Baron & Greene, 1996; Desvousges et al., 1993; Kahneman & Knetsch, 1992), charitable giving (Västfjäll and Slovic, 2020, see also Chapter 4 of this thesis), or in the reaction to mass suffering and genocides (Cameron & Payne, 2011; Dickert et al., 2015; Slovic & Västfjäll, 2010a). Like narrow bracketing, scope insensitivity would suggest that people are not sufficiently sensitive to the amount of time that is lost by extinction (or the number of future choices), and would therefore likely lead to deviations from optimal decision making under extinction risk.

In Experiment 3b, we investigated how people adjust their choices as the maximum trial number changes. We decided on the maximum trial number in the different conditions based on three considerations. First, we chose maximum trial num-

bers that imply a sufficient difference in optimal strategies to maximise the chance of finding an effect of maximum trial number if it is present. Second, we tried to avoid extreme cases where the optimal solution implies always or never choosing the risky option. Third, we wanted to retain a maximum trial number of 100 in one of the two conditions to allow a direct replication of Experiment 1.

Based on these considerations, we chose 60 and 100 as our maximum trial numbers for comparison. As the maximum remaining trials should mostly have an effect in the Lose and Keep conditions, where earnings from a large number of future trials can be foregone upon extinction, this experiment focuses on those two conditions. This resulted in the following four groups, with the optimal number of risky choices indicated in brackets:

1. 60 maximum trials – Lose condition (optimal strategy $\approx 14/60$ risky choices)
2. 100 maximum trials – Lose condition (optimal strategy $\approx 10/100$ risky choices)
3. 60 maximum trials – Keep condition (optimal strategy $44/60$ risky choices)
4. 100 maximum trials – Keep condition (optimal strategy $44/100$ risky choices)

Even though the difference between 60 and 100 trials might appear small, the implied difference in the proportion of risky choices is substantial. In the Lose condition, participants should choose the risky option more than twice as often for 60 compared to 100 maximum trials (23% vs. 10% of the time), and in the Keep condition, a bit less than twice as often (73% vs. 44% of the time).

In addition to testing the effect of maximum trial number, Experiment 3b also aimed to replicate the sensitivity to the core task characteristics found in the previous experiments (in particular, Experiment 1, which compared the same conditions but only for 100 maximum trials).

Overall, Experiment 3b tested one set of predictions and one research question:

1. **Sensitivity to Extinction Condition:** We expected the same effects as observed in Experiment 1. The proportion of risky choices in the Keep conditions will be higher than in the Lose conditions. Additionally, we expect the

same interaction between Extinction Condition and trial number: Participants in the Keep conditions will increase the proportion of risky choices across trials, whereas participants in the Lose conditions will not. (preregistered, see: <https://osf.io/wesxa>)

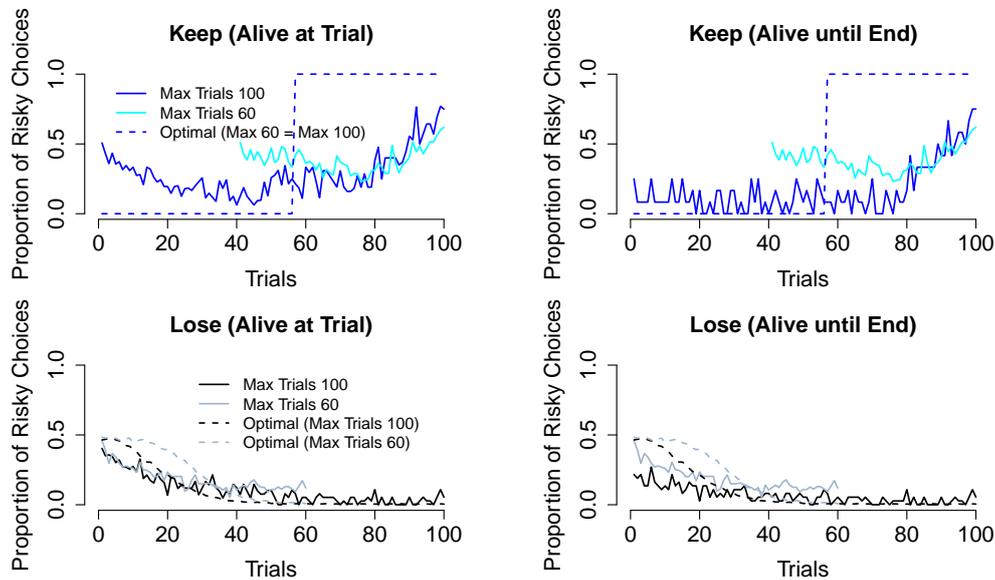
2. **Sensitivity to Maximum Number of Trials:** Are participants playing more risky when the maximum number of trials is lower? (not preregistered)¹¹

Participants. Participants were paid £1.20 to participate in an 8-minute study. On top of that base payment, participants in the Lose condition earned an average bonus of £0.52 IQR [£0.00, £0.82] and in the Keep condition £0.77 IQR [£0.22, £1.19]. The study was approved by the ethics chair of UCL's Department of Experimental Psychology (EP/2021/005). Initially, 393 participants completed the study. We excluded participants based on four different comprehension/attention checks. Four participants indicated the wrong number of trials, 14 participants did not indicate correctly that the piggy bank in the top right corner showed their current bonus earnings, 18 participants did not remember the possible payoffs in the experiment correctly, and 65 participants indicated that they did not understand the nature of the extinction event. Some participants failed multiple of these checks. After these exclusions, we obtained a final sample of 305 participants. The mean age of participants was 40.35 years (SD = 13.03). 158 participants were female and 147 male. 142 participants were in the Lose condition (85 in the 60-trial condition and 57 in the 100-trial condition) and 163 participants in the Keep condition (96 in the 60-trial condition and 67 in the 100-trial condition).

Design, Materials and Procedure. The method was the same as in Experiment 1 aside from the addition of the maximum trial number manipulation (60 vs. 100), and the counterbalancing of the position of the two lotteries (as in Experiments 2 & 3a). This resulted in a 2×2 (Extinction Condition \times Maximum Trials) between-participants design.

¹¹For this analysis, we deviate from the preregistration with an analysis that maps better onto our theoretical question. We report the preregistered analysis in a footnote in the Results section.

Figure 5.10: Risky Choices for Each Maximum Trial Number and Extinction Condition for Experiment 3b



Note. In the Keep condition, the 60-trial condition is overlaid on the 100-trial condition in such a way that they both show the same end point. Therefore, the lines for the optimal solutions exactly overlay for 60 vs. 100 Maximum Trials. The optimal solution for the Lose condition is based on simulating 5000 participants, which follow the dynamic solution (Equations 5.4–5.7). The decision is executed based on a softmax decision function, as otherwise tiny differences in expected value in the first choices would lead to deterministic switching between 0 and 1. We use a very low temperature (0.02) so that the solution is very similar to a deterministic decision rule. The same qualitative pattern arises for a range of temperature values. For a version of this figure that shows model predictions from the fitted mixed effects model along with 95% confidence bands see <https://osf.io/pz7hg>.

Results. As in Experiment 1, participants' choices differed between the Keep and Lose conditions. In line with the optimal strategies, participants made more risky choices at the end of the experiment in the Keep condition, whereas they made more risky choices at the beginning of the experiment in the Lose condition (Figure 5.10). Regarding the influence of Maximum Trials, we found that in the Keep conditions, participants increased their risky choices a similar number of trials from the end of the experiment, regardless of whether they would see a maximum of 60 or 100 trials (Figure 5.10, top panels, in which the 60-trial condition has been plotted according to the distance from the final trial). In the Lose conditions, participants played a similar proportion of risky choices in the first 60 trials, again regardless of whether they would see a maximum of 60 or 100 trials (Figure 5.10, bottom panels).

We first tested participants' sensitivity to the Extinction Condition using a logistic mixed effects model with effects of trial number (linear & quadratic effect), Extinction Condition, Maximum Trials, their two-way interactions, and their three-way interaction. We standardised the trial number by subtracting the midpoint within each condition.¹² As predicted, and as in Experiment 1, we found a significant difference in the estimated average riskiness, $z = 9.04, p < .001$, and a significant interaction between Extinction Condition and trial number (linear effect), $\chi^2(1) = 472.92, p < .001$, with a positive marginal slope in the Keep conditions, $b = 1.74, 95\% \text{ CI } [1.47, 2.00]$, and a negative marginal slope in the Lose conditions, $b = -2.17, 95\% \text{ CI } [-2.46, -1.89]$.¹³

The visual pattern from Figure 5.10 suggests that in the Keep conditions, participants played the same proportion of risky choices with the same distance from the *end* of the experiment, whereas in the Lose conditions, participants played the same proportion of risky choices at the same distance from the *beginning* of the experiment. We therefore estimated two additional mixed effects models, one for each Extinction Condition, that compares the estimated average riskiness of the conditions with 60 maximum trials to the estimated average riskiness in the first 60 trials (Lose) and the last 60 trials (Keep) of the conditions with 100 maximum trials. Consistent with the visual impression from Figure 5.10, this indicated no evidence for an effect of Maximum Trial Number in the Keep conditions, $z = 0.65, p = .518$, nor in the Lose conditions, $z = 1.51, p = .131$.¹⁴

¹²Specifically, for the 60-trial conditions: $trialNr_{\text{stan}} = (trialNr - 30.5)$ and the 100-trial conditions: $trialNr_{\text{stan}} = (trialNr - 50.5)$ (Note that the mean of trials ranging from 1 to 100 is 50.5 not 50).

¹³We further found a significant interaction between trial number (quadratic term) and condition $\chi^2(1) = 23.24$; Keep: $b = 6.25, 95\% \text{ CI } [5.37, 7.13]$; Lose: $b = 2.75, 95\% \text{ CI } [1.78, 3.72]$. These results were based on a random intercepts only model. The maximal model, which showed convergence issues (Hessians with negative Eigenvalues) lead to the same conclusions. Further, in this experiment, we preregistered only a linear effect of trial number. Removing the quadratic effect of trial number lead to the same conclusions.

¹⁴Testing for an effect of maximum trial number using the first GLMM (i.e., the one used to test for the main effect of Extinction Condition), shows evidence for an effect of Maximum Trial Number in the Lose condition, $z = 2.34, p = .020$, but not in the Keep condition, $z = 0.15, p = 0.883$. The reason is that this analysis does not compare matching trials but rather compares the estimated average riskiness across 60 trials in the 60-trial condition to the estimated average riskiness across 100 trials in the 100-trial condition.

Computational Modelling Results. The models again showed good convergence (all $\hat{R} < 1.01$). As in Experiment 1, we found a significant difference in strategies employed between Extinction Conditions, $\chi^2 = 33.31, p < .001$. More participants followed the risky strategy and the safe to risky switch strategy in the Keep than in the Lose conditions. We found no evidence for an impact of the Maximum Trial Number on the employed strategy (Lose: $\chi^2 = 7.39, p = .191$, Keep: $\chi^2 = 3.18, p = .682$; see Appendix C.8.4 for visualisations of the strategies as a function of Extinction Condition and Maximum Trial Number).

Discussion. Experiment 3b replicated the findings of Experiment 1, where participants responded to the differences between the extinction conditions in ways that are qualitatively in line with the optimal strategies (i.e., they increased risky choices towards the end of the experiment in the Keep but not in the Lose conditions). Regarding the impact of the maximum trial number, our results suggest that in the Lose conditions, participants' choices depended on the number of trials already played (independent of the maximum trial number), while in the Keep conditions, their choices depended on the number of trials remaining (again independent of the maximum trial number). These results are consistent with the hypothesis that choice bracketing and scope insensitivity may lead people to be insensitive to the maximum trial number.

However, in addition to this insensitivity, participants' behaviour also showed a surprisingly adaptive element in the Keep conditions, which diminishes deviation from optimality: they switched at a constant distance from the last trial (rather than, for instance, after a constant proportion of trials). Therefore, insensitivity to trial number only leads to a deviation from optimality in the Lose conditions, where the number of risky choices changes as the number of trials reduces. Participants in this condition played somewhat too safely when the maximum trial number is reduced (since the optimal strategy implies more risky choices for fewer trials, red dotted line in Figure 5.10). In the Keep conditions, insensitivity to trial number leads to well-calibrated behaviour, because people switch on average with the same distance from the end, a behaviour that is aligned with the optimal strategy (black

line in Figure 5.10).

5.1.5 General Discussion

This chapter aimed to advance our understanding of decision making in the face of irredeemable losses—namely, extinction risks. We proposed three definitions of extinction or catastrophic risk for study in a decision making task: (1) Lose, where all endowment is wiped out upon extinction and the participant cannot earn any payoffs for trials after the extinction event; (2) Keep, where the endowment is retained upon extinction, but the participant cannot earn any payoffs for trials after the extinction event; and (3) Reset, where the endowment is wiped out, but the participant can continue playing. We derived optimal strategies for these three definitions and operationalised them in a new experimental paradigm, the Extinction Gambling Task (EGT).

Leveraging the EGT and the computational models developed for it, we were able to obtain a nuanced understanding of people’s choices in the task, which highlights both strengths and weaknesses in human decision making under extinction risk. On the positive side, the strategies employed by participants indicated a general qualitative understanding of the differences between extinction conditions. However, our studies also indicate several deficits: consistent with opportunity cost neglect (Frederick et al., 2009), we found that people are more in line with the optimal strategy and less risk-seeking when their endowment is at stake than when the opportunity to keep playing is at stake (Experiment 2); in line with research on loss chasing (Gainsbury et al., 2014; Lister et al., 2016), we found that introducing losses leads to excessive risk-taking (Experiment 3a); in line with research on scope insensitivity (Desvousges et al., 1993) and choice bracketing (Read et al., 2000), we found that participants in the Lose condition are not appropriately sensitive to the maximum trial number (Experiment 3b).

A pertinent question concerns how the results from the ‘small world’ (Savage, 1972, pp. 82-91) of our task relate to decision making about extinction risk in the real world. There are several real-world examples suggesting that people might be dealing with extinction risks rather poorly (e.g., Wiener, 2016). For instance,

despite the recent pandemic and the possibility of another more deadly pandemic during our lifetimes (Marani et al., 2021), governments' investment in pandemic preparedness is relatively small compared to other areas of spending (Clark et al., 2022; Michael & Mark, 2024). Some of our findings are consistent with poor decision making about extinction risks. Specifically, people's choices in our task are affected by irrelevant factors (e.g., the introduction of losses and endowment in Experiment 3a), and sometimes participants are not sufficiently sensitive to other factors that should affect decision making (e.g., the total number of trials in the Lose condition of Experiment 3b). Some of the phenomena that we document have also been found in more realistic (albeit less controlled) settings (e.g., Coleman et al., 2023; Slovic & Weber, 2013). These observations suggest that the task may indeed be appropriate for isolating psychological factors that affect real-world decision making.

However, we do not only find limitations in human decision making. Participants were relatively good in terms of the qualitative strategies that they employed and they were sensitive to the differences between extinction conditions. To the extent that people make reasonably good decisions in our task but bad decisions in a corresponding real-world scenario, this would suggest shortcomings in real-world decisions other than an inability to understand the nature of the risk. In the next paragraphs, we outline several modifications of our task that may further the understanding of decision making under extinction risk.

Where Next? The Future of the Extinction Gambling Task

The current experiments, and the current version of the EGT, provide a starting point for research into decision making under extinction risk. The task offers a flexible framework that can be extended to address a wide range of theoretical and applied questions in both individual and collective decision making. Here, we outline some directions for future research using the EGT.

Answering Theoretical and Practical Questions in Different Risk-Taking Domains. Within the EGT, decisions under extinction risk are investigated via mon-

etary gambles. This has a variety of advantages, in particular the ability to derive optimal strategies and compare participants' behaviour to them. However, our results should therefore be interpreted cautiously as speaking most directly to decisions under extinction risk within this domain. The observed choice patterns may be substantially different when examining different domains, types of gambles, or increasing the stakes (Blais & Weber, 2006; Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Weber et al., 2002). The task could therefore be modified to investigate these differences. For instance, one could imagine a version with a realistic cover story related to pandemic response, where not locking down an area would correspond to the risky option, which can result in a large-scale outbreak or a global pandemic (corresponding to an extinction event), but in the more likely event that an outbreak does not happen, may lead to more economic activity and fewer people affected by control measures. Such modifications would allow bridging the gap between the current financial gambles and decision making under extinction risk in specific real-world domains. In general, decisions under extinction risk can be made increasingly more realistic by giving up internal validity for increased external validity: first, one could introduce realistic cover stories, while retaining the information about monetary outcomes (so it is still possible to calculate optimal strategies); second, one could create a purely hypothetical task with a realistic cover story but no monetary element; third, one could investigate risky-choices in the real world and see whether age-related changes reflect patterns of optimal strategies in our task.

Future work might additionally explore changes of the task structure, for instance by introducing multiplicative rather than additive payoffs (e.g., one's wealth grows by a constant percentage every trial if one does not go extinct), variable risk over trials, situations where an extinction event is possible for both lotteries, or allowing players to invest some money in order to reduce the risk. This creates further links to research in fields adjacent to psychology, particularly optimal foraging theory (Stephens & Krebs, 1986) and economics, and allows modelling behaviours such as saving for retirement (AXA Pensions, 2024; Vanguard, 2021), saving in

preparation for potential unexpected costs (Wang-Ly & Newell, 2024), and investment portfolio choice. Indeed, we can already see links between the present findings and commonly observed effects in these domains. For example, in investment portfolio choice, the finding that wealthier individuals typically invest in more risky assets (Guiso & Paiella, 2008; Peress, 2004) is reminiscent of the increase in riskiness during the task in the Keep condition (though important differences remain in terms of the implied optimal strategies and the observed behaviour, for instance the Keep condition has an optimal single switch strategy, which is not applicable in portfolio choice).

Further, the EGT could be modified to test to what extent people follow through on their planned behaviour and how this affects decision making about extinction risks. In the experiments reported here, participants engaged with the EGT dynamically (i.e., they experienced the outcomes of their choices in each trial before making the decision for the next trial). We argue that this dynamic implementation mimics the way people make most decisions under extinction risk in the real world. Consequently, the EGT should be prone to some of the same biases that operate in real-world decision making under extinction risk, in particular, survivorship or observer selection effects (e.g., when alive, one has never experienced a fatal event). Previous work has investigated decision making about extreme risks in static setups (i.e., all choices are specified in advance; Perfors & Van Dam, 2018) or a combination of static and dynamic setups (Crosetto & Filippin, 2013). Moving forward, it would be interesting to leverage the dynamic aspect of the EGT to study discrepancies between planned and on-the-spot actions, also known as dynamic inconsistencies (e.g., Barkan & Busemeyer, 1999; Cubitt et al., 1998; Hotaling & Kellen, 2022).

Finally, all elements of the task were known to participants in our experiments (maximum trial number, payoffs and probabilities, and current endowment), as this information is required to calculate the optimal strategies. Future work could investigate what happens when information about these factors is reduced. Regarding the endowment, removing the display may reduce participants' ability to take their

accumulated earnings into account, which could increase risk-taking, whereas not knowing the probabilities and outcomes could reduce risk-taking, similarly to ambiguity aversion (Ellsberg, 1961). The effect of not knowing the maximum trial number is more difficult to predict and would likely depend on people's prior beliefs.

Studying Collective and Social Decisions. Many of the extinction risks that people are exposed to are collective rather than individual risks. For instance, when taking measures to mitigate climate change, if one country does not reduce its emissions but other countries do, this country could 'free ride' and enjoy the benefits of mitigated climate change without the cost of reducing its own emissions. Public goods tasks that study cooperation in the context of preserving resources for the next generation (Hauser et al., 2014; Jacquet et al., 2013) find that, absent any interventions to improve cooperation, defection is usually too high to preserve the public good. In order to investigate collective decisions under extinction risk, we are currently working on adapting the task into a public goods game, where players play in groups of five. In this type of task, choosing the risky option constitutes a form of defection as the player who plays risky accumulates all the payoffs from the risky decision themselves, but if they draw the extinction option, the whole group will go extinct. By comparing risk-taking in the individual-level task to risk-taking in the collective task, we can quantify to what extent non-optimal decision making about extinction risk is driven by cooperation problems. Additionally, a collective task will allow us to test interventions to improve collective decision making, such as median voting (Hauser et al., 2014) or increased communication between participants.

Another way in which the EGT could be modified to account for social interactions is by incorporating competition between different players, which would be similar to some existing dice games, and has been shown to affect risky choices in previous work (Schulze et al., 2015).¹⁵ A competitive setup would introduce additional game-theoretic elements as the optimal strategy in this type of game also

¹⁵For instance, in 'Pass the Pigs', two players need to roll a set of two-pigs and are scored depending on which side the pig falls on. For a mathematical treatment, see <https://www.youtube.com/watch?v=ULhRLGzoXQ0>.

depends on what the other players would choose. For example, if the other players are very risky and likely to go extinct, one can win easily by being safe; on the other hand, if the other players follow a safer strategy, it becomes optimal to play more risky to increase variance and have a chance of achieving a score that is higher than each of the other players. These competitive dynamics may play out in real-world decisions, such as when multiple companies compete to develop a potentially dangerous technology. The competitive setup would allow us to test to what extent risk-taking is increased compared to the individual level game and allow testing interventions to reduce risk-taking (e.g., communication and contracts between the players).

Can the Task Be Used to Measure Individual Differences? The EGT as implemented in this chapter was not designed for measuring individual differences and we caution against directly using the task as individual difference measure, as the current one-shot setup with a single block and one extinction event likely has low reliability. Reliability could, however, be increased by chaining the current task multiple times in a row. The current duration of the main part of the task is around 6 minutes, so chaining the task would be feasible in terms of duration and participant engagement. However, creating a multi-block version would distort the nature of the extinction event, as it would not wipe out all earnings, but only the earnings for this block. It would also allow participants to learn from their experiences of extinction events. Both of these properties are in conflict with the definition of extinction events we outlined in the introduction. Future research investigating the potential differences between repeated and one-shot designs is therefore necessary before versions of the task that are suitable for measuring individual differences can be employed.

These extensions of the EGT go hand in hand with extensions to the modelling of the choice data on an individual level. The (dependent) mixture model employed here can be considered a descriptive or measurement model that allows us to classify the strategies based on the choice data, but the parameters of the model do not have a straightforward psychological interpretation. In future work, it would be interesting

to consider process models, for instance, reinforcement learning models (Watkins, 1989; Watkins and Dayan, 1992, especially so for modifications of the task without known probabilities and/or payoffs), or risky choice models (De Palma et al., 2008; Krueger et al., 2024; Tversky & Kahneman, 1992). However, when applying risky choice models, additional questions arise, such as whether risk aversion should operate on a per-trial basis or over the whole set of choices in the task. To address these questions, one could compare responses and parameter estimates between the EGT and decisions in a risky-choice task not involving extinction risk.

Closing Remarks

Until now, little psychological research has investigated decisions under extinction risk. This chapter introduced the Extinction Gambling Task (EGT) as the first paradigm designed to study these decisions. While our findings offer valuable psychological insights, the primary aim of this work was to lay a robust foundation for future research in this area. Given the applied relevance of decisions under extinction risk, it is important to consider the usefulness of a laboratory task like the EGT. Some researchers have argued that much of decision making research has had very limited real-world impact, owing to its emphasis on task designs tailored to testing formal models in artificial settings rather than examining truly ‘consequential decisions’ (Weiss & Shanteau, 2021). While the study of consequential decisions is indeed vital, we contend that meaningful laboratory tasks can serve this goal, provided they capture features that are genuinely relevant to real-world decision contexts (for relevant discussions, see Fiedler, 2018; Garcia-Marques & Ferreira, 2018). The EGT has been developed primarily with this goal in mind. The computational modelling conducted here was designed to understand the nuances in the observed choice data, rather than for testing different formal models. This enables researchers to observe individual choice strategies in scenarios with meaningful features that isolate key properties of decisions under extinction risk, such as repeated choice opportunities, wealth accumulation, and the possibility of irredeemable outcomes. Given how these features map onto real-world scenarios, we are optimistic

about the EGT's potential for facilitating scientific insight. We hope that our work and the development of the EGT will help to bring much-needed research attention to a significant domain that has been largely ignored.

Reference for Journal Article: Maier, M., Harris, A. J. L., Kellen, D., & Singmann, H. (in press). Decision Making Under Extinction Risk. *Cognitive Psychology*. <https://doi.org/10.31234/osf.io/qjd37>

Data, Code, and Preregistration Availability: All data and code are available at <https://osf.io/qhkw6/>.

Supplemental Information: Supplementary Materials are available at <https://osf.io/qhkw6/>.

Author contributions: All authors were involved in the conceptualization and design of the experiments. M.M. created the materials and programmed the experiments with feedback from the other authors. M.M. conducted data analysis. M.M. with feedback from H.S. M.M wrote the first draft of the manuscript. All authors contributed to writing, reviewing, and editing the manuscript.

5.2 Collective Decision Making Under Extinction Risk (Stage 1: Registered Report with Pilot Data)

As outlined in the discussion of the previous section, people not only face extinction risks in individual decisions, but humanity also faces *collective* extinction risks – such as runaway climate change (Kemp et al., 2022), risks from artificial intelligence (Christian, 2021; Russell, 2019), bio-terrorism and pandemics (Millett & Snyder-Beattie, 2017a, 2017b), and nuclear war (Ellsberg, 2017; Perry & Collina, 2020). In this section, we therefore generalise the Extinction Gambling Task to study collective decision making under extinction risk.

Appropriate risk-taking in the face of collective extinction risks depends on both individual and collective reasoning and decision making: individuals need to understand the risks and know how to act to reduce them, and at the collective level, coordination and cooperation between individual actors are needed to keep the overall risk within appropriate limits. Society's response to risks such as climate change

highlights the difficulties that arise on both of these levels. At the individual level, it is difficult to convince the public to take such an abstract and long-term threat seriously (Weber, 2010); at the collective level, governments face complex coordination problems during international negotiations (Bodansky, 2001), and individual states may be able to ‘free-ride’ if other states reduce their emissions and they do not. These challenges also shape mainstream interventions, with bottom-up efforts to educate the public (Monroe et al., 2019) and top-down institutional frameworks like the Conference of the Parties (COP) and climate change agreements (Falkner, 2016) to coordinate the global response.

Studying collective decisions under extinction risk, therefore, requires taking factors of both individual and collective decision making into account. Investigating only individual-level limitations fails to take into account the public goods structure of the problem. On the other hand, public goods games construe collective decisions under extinction risk as a cooperation problem from the outset, without considering limitations in individual-level reasoning. To take a more integrated approach, this section compares individual and collective decision making under extinction risk using a novel behavioural experiment which involves choices under extinction risk in both individual and group conditions. In the group game, we further include interventions to address (1) individual-level reasoning limitations through providing participants with information about the optimal strategy (Coleman et al., 2023), and (2) coordination and cooperation challenges through a voting institution (Hauser et al., 2014).

Shortcomings in people’s decision making under extinction risk outlined in the first half of this chapter are likely compounded by the need for cooperation and coordination to manage extinction risks. Avoiding extinction risks is a public goods dilemma (Dreber & Nowak, 2008; Hauser et al., 2014; Milinski et al., 2008). The public good, in this case, would be to preserve a low collective risk of extinction. For any individual actor, it might be worthwhile – from a self-interested perspective – to increase the overall risk of extinction for everyone involved if this results in a personal gain (e.g., a country not investing in reducing CO₂ emissions and still

benefiting from other countries doing so). However, collectively, everyone is worse off when the risk of extinction is increased. Conversely, if one actor invests resources to reduce extinction risk, the benefit is shared by everyone. Therefore, a set of self-interested utility-maximising agents could end up in a state of high risk, even though they would collectively be better off if risks were lower (see Appendix C.9, which generalizes the individual strategies to the collective task). Previous public goods games have investigated collective decision making about risks in the form of ‘collective risk social dilemmas’ (Milinski et al., 2008) or ‘intergenerational goods games’ (Hauser et al., 2014). In these games, participants usually fail to manage the risk by contributing less to preserving the public good than would be collectively optimal; however, no previous work has compared collective and individual decision making within the same paradigm.

Shortcomings on the individual and collective level also point to different types of interventions to reduce risk-taking. At the individual level, educational interventions may increase awareness of risks and willingness to act in order to mitigate them. For instance, various educational interventions have been developed to increase awareness and understanding of the challenges posed by climate change (e.g., Leal Filho et al., 2021; Monroe et al., 2019). Other work has tried teaching expected value calculation and increased scope sensitivity to increase willingness to engage in extinction risk mitigation more generally, though these interventions showed only limited effectiveness (Coleman et al., 2023). At the collective level, institutional frameworks aim to increase coordination and cooperation. Binding agreements or shared norms established through international institutions can provide a framework for long-term cooperation on global risks, such as the Paris Agreement for climate change (Bodansky, 2001; Falkner, 2016). One promising approach is the implementation of binding voting mechanisms or democratic decision making procedures, which have been shown to increase cooperation in public goods dilemmas (Bernard et al., 2013; Hauser et al., 2014; Walker et al., 2000).

In this section, we therefore investigate individual and collective decision making under extinction risk in a joint paradigm that allows us to estimate to what

extent cooperation and coordination problems impact decision making under extinction risk relative to limitations in individual-level decision making. Further, we investigate interventions to reduce risk-taking aimed at improving decision making targeted at improving individual-level reasoning versus improving cooperation.

To study these questions, we extend the task introduced in this chapter to study collective decision making under extinction risk. As before, participants play a series of risky choices, where they decide between two lotteries: a safe lottery that has no risk of extinction and a risky lottery that has a small risk of extinction but also a chance of a higher payoff. If participants choose the risky lottery and draw the extinction outcome, they receive no bonus payments for the whole experiment.

However, in the collective version of the task, participants do not play alone but instead play in groups of five. Each participant decides between the safe and risky lotteries. However, depending on the outcome, their choice affects everyone in the group: they retain bonus payments from their own choices, but if any one participant draws the extinction event, the whole group receives no bonus payment for the whole experiment. We contrast responses in the collective task to responses in the individual task.

In the individual task, one individual faces a sequence of 100 choices, while in the collective task, five individuals face a sequence of 20 choices each (i.e., 100 choices in total). The optimal *collective* bonus payoff and optimal strategy in this version, where 5 individuals face 20 choices each, is extremely similar to the optimal individual bonus payoff and strategy in the version where one individual faces 100 choices.¹⁶ However, the individually optimal solution in the collective game is considerably more risky than the individually optimal solution in the individual game (see Appendix C.9). Therefore, we can test how much cooperation and co-

¹⁶A slight difference arises because in the dynamic optimal solution to the task, the individual has more information than the group (as individuals learn about outcomes after every choice rather than in batches of five), but this affects payoffs of the optimal strategies by less than 1p (see Appendix C.9). Another difference arises because the group needs to coordinate in order to make the optimal number of risky choices (e.g., even if the optimal strategy was known and all players were altruistic, it would be difficult to, for instance, coordinate on playing 8 out of 100 risky choices, since the players cannot communicate); Appendix C.9 shows that taking these potential coordination problems into account does not affect the optimal payoffs considerably.

ordination problems affect performance in the task by comparing decisions in the individual task to decisions in the collective task.

We further test the effectiveness of two types of interventions designed to reduce risk-taking under extinction risk, one aimed at improving participants' individual decision making abilities and one aimed at improving cooperation. As an intervention to improve individual-level reasoning about the optimal strategy, we introduce a display that informs people how many risky choices they still need to play to achieve optimal collective expected value in the game. In light of findings showing that it is often difficult to improve decision making with individual level 'debiasing' interventions (e.g., Coleman et al., 2023; Meyer & Frederick, 2023; Singmann et al., 2024), we chose this relatively strong intervention to ensure that we would be able to detect an effect of individual level reasoning interventions if these can play a role at all.

As an intervention to improve cooperation, we introduce a median voting condition, where each participant votes on the number of risky choices that should be played as a group, and the median of the vote is then executed as the number of risky choices in that round. This intervention has been shown to be successful in improving cooperation in other types of public goods games (Bernard et al., 2013; Hauser et al., 2014) and is therefore a promising contender for reducing risk-taking in light of extinction risk.

The two types of interventions are similar to the two approaches that can be taken to reduce extinction risk: better education and information versus top-down institutional frameworks. Since the two types of interventions are reminiscent of approaches to address societal risks, both interventions are applied to the collective task. Finally, we also include a combined intervention which includes both median voting and information about the optimal strategy.

Overall, our study has five experimental conditions (individual, collective, collective + information, median voting, & median voting + information). By comparing these conditions, we can isolate key factors that shape decision making under extinction risk and consequently address several questions related to why people

may neglect extinction risks and interventions to reduce risk-taking (see also Table 5.1 for a more detailed overview):

- **RQ1:** Is risk-taking in the context of extinction risks partially driven by cooperation problems? (individual vs. collective)
- **RQ2:** Does information about the optimal strategy reduce risk-taking in collective decisions under extinction risks? (collective vs. collective + information)
- **RQ3:** Does median voting reduce risk-taking in the context of extinction risks? (collective vs. median voting)
- **RQ4:** Does a combination of median voting and information about the optimal strategy reduce risk-taking in the context of extinction risks more than each of them in isolation? (median voting vs. median voting + information & collective + information vs. median voting + information).

In this chapter, I present a pilot study, which includes all relevant experimental conditions to address the research questions above (albeit with a smaller sample size). This pilot data was used in a submission of the collective extinction task as a registered report.¹⁷ Therefore, this chapter includes both information about the pilot study (i.e., the methods section includes information about the pilot sample) as well as a separate sampling plan section with the intended sample size for the main study. Because the sample size for the pilot is relatively large for a pilot study ($N = 294$), it already reveals some interesting patterns, which I hope will make it of interest to the reader.

¹⁷For more details on registered reports, see <https://www.nature.com/naturebehavior/submission-guidelines/registeredreports>.

Table 5.1: Design Table

Question	Hypothesis	Analysis Plan	Interpretation
Is risk-taking in the context of extinction risks partially driven by cooperation problems?	The estimated average riskiness in the collective condition is higher than in the individual condition.*	Logistic mixed effects model, comparing the estimated average riskiness between these two conditions.	If estimated average riskiness is significantly higher in the collective condition than in the individual condition, we conclude that risk-taking is partially driven by coordination and cooperation problems. Otherwise, we conclude that there is no evidence that risk-taking is driven by coordination and cooperation problems.
Does information about the optimal strategy reduce risk-taking in collective decisions under extinction risks?	The estimated average riskiness in the collective + information condition is lower than in the collective condition.	Logistic mixed effects model, comparing the estimated average riskiness between these two conditions.	If estimated average riskiness is significantly lower in the collective + information condition than in the collective condition, we conclude that providing information about the optimal strategy reduces risk taking. Otherwise, we conclude that there is no evidence that providing information about the optimal strategy reduces risk taking.
Does median voting reduce risk-taking in the context of extinction risks?	The estimated average riskiness in the collective + median voting condition is lower than in the collective condition.	Logistic mixed effects model, comparing the estimated average riskiness between these two conditions.	If estimated average riskiness is significantly lower in the collective + median voting condition than in the collective condition, we conclude that median voting reduces risk taking. Otherwise, we conclude that there is no evidence that median voting reduces risk taking.
Does a combination of median voting and information about the optimal strategy reduce risk-taking in the context of extinction risks more than each of them in isolation?	The estimated average riskiness in the combined intervention condition is lower than in either of the other intervention conditions.	Logistic mixed effects model, comparing the estimated average riskiness of the combined condition to the median voting only and the optimal information-only condition.	If both tests are significant in the predicted direction, we conclude that a combined intervention is more effective; if only one is significant, a combined intervention is only better than one of the two interventions; if none are significant, we conclude that there is no evidence a combined intervention reduces risk-taking over each intervention in isolation.

Note. *For more details on how we define ‘estimated average riskiness’ and the mixed effects model specification and power, see the Method section as well as Appendix C.10).

5.2.1 Method

Ethics Information

The research complies with all relevant ethical regulations and has received ethics approval by the Departmental Ethics review board of UCL Experimental Psychology (EP/2021/005). We obtained informed consent from all participants. Participants were paid a base payment of £2.25 for participation in a 15 minute study. In addition to that base payment, participants received a bonus payment depending on their performance.

Design

Conditions. Participants were randomly assigned to five experimental conditions between-subjects:

1. **Individual:** One participant played a sequence of 100 trials in a single-player game.
2. **Collective:** 5 participants played a sequence of 20 trials in a collective goods game.
3. **Collective + Information:** This collective condition also provided participants with information about the optimal number of risky choices they should play as a group.
4. **Median Voting:** In this collective condition, each participant indicated how many risky choices the group should make in total in the current round, and the median vote was then executed. It was then randomly assigned which players played the safe and which played the risky choices.
5. **Median Voting + Information:** This condition included both the median voting and the information manipulations.

Payoffs and Probabilities. Each game had a total of 100 trials. For the individual game, this implies a single player playing 100 trials, whereas for the collective game with 5 players, this implies 20 trials per player.

In the individual game, participants needed to decide between the following two lotteries in each trial:

- Risky Lottery:
 - 47.5% chance of winning 0p,
 - 47.5% chance of winning 10p,
 - 5% chance of extinction (i.e., no bonus payoff at all for the whole experiment).

- Safe Lottery:
 - 50% chance of winning 0p,
 - 50% chance of winning 1p.

We chose this setup as it creates some benefit of going with the risky option on a per-trial basis (10p vs. 1p), similar to real-world decisions where there is usually some (short-term) benefit associated with not mitigating extinction risks (e.g., cost-savings). Appendix C.9 shows optimal strategies for this game setup. This indicates that the optimal strategy implies choosing ≈ 8 risky choices and otherwise safe choices (the actual optimal policy is dynamic which means that the precise number of risky choices depends on luck). However, if a participant or group plays too safe, this results in only a modest reduction in expected bonus payoff compared to following the optimal strategy (the expected value of the all-safe strategy would be 50p vs. 58p for the optimal). On the other hand, the expected earnings decline rapidly as the number of risky choices increases (the expected value of playing all risky choices is only 2.96p). This reflects a property that is likely shared with real-world extinction risks: if in doubt, it is better to err on the side of being too safe than on the side of being too risky.

For the collective-level game, if one participant drew the extinction event, everyone went extinct. Further, all per-trial payoffs were multiplied by a factor of five. In other words, the risky option had a 47.5% chance of resulting in a payoff of 50p (instead of 10p), and the safe option had a 50% chance of resulting in a payoff of 5p (instead of 1p). This ensured that if the group follows the collectively optimal strategy, each individual player receives, in expectation, the same bonus payoff as a player participating in the individual game. Appendix C.9 shows that the group optimal strategy and the individually optimal strategy are almost the same between the collective and the individual game.

Materials & Procedure. Participants were recruited from the Prolific platform and participant pool (<https://www.prolific.com>). We screened participants to only include those who are based in the United Kingdom, have English as a first language, and an approval rating of more than 95%. Participants received a link that took them to an online game built using the oTree platform (version 5.0.0) (Chen et al., 2016). All code for the game is included in the GitHub repository https://github.com/chasmani/extinction_group_game. Before starting the game, participants completed a detailed consent form, and they had the option to leave the experiment at this point.

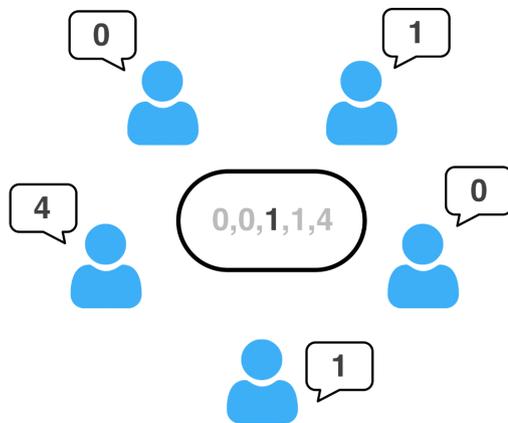
Participants first received detailed task instructions, including the maximum number of trials, the lotteries, and associated probabilities, and what happens when they draw the extinction event. We also included the following two comprehension checks: ‘What happens if you choose the risky lottery and you draw the extinction option?’ (correct answer: ‘I will lose all my bonus money, and I cannot get any more bonus payments for future rounds.’) and ‘How many rounds will you play in the experiment?’ (correct answer collective conditions ‘20 rounds’, correct answer individual conditions: ‘100 rounds’).

In the group extinction condition, participants then further read further information about the group decision making process, in particular, that everyone receives the payoffs of their own choices, but if one player goes extinct, everyone goes extinct. They also needed to answer the following comprehension check: ‘What

happens if one of the players draws the extinction outcome in the risky lottery? (correct answer: ‘The entire group will go extinct, and lose all the bonus.’).

For the median voting condition, this page also included information about how the voting process works, including two examples of how participants’ votes lead to the group’s risky decisions (see Figure 5.11). They were further asked the following comprehension check ‘How is the group decision made for how many players will play the risky lottery?’ (correct answer: ‘The group decision is the middle value of the ordered votes.’)

Figure 5.11: Illustration of the Median Voting Condition as Shown to Participants



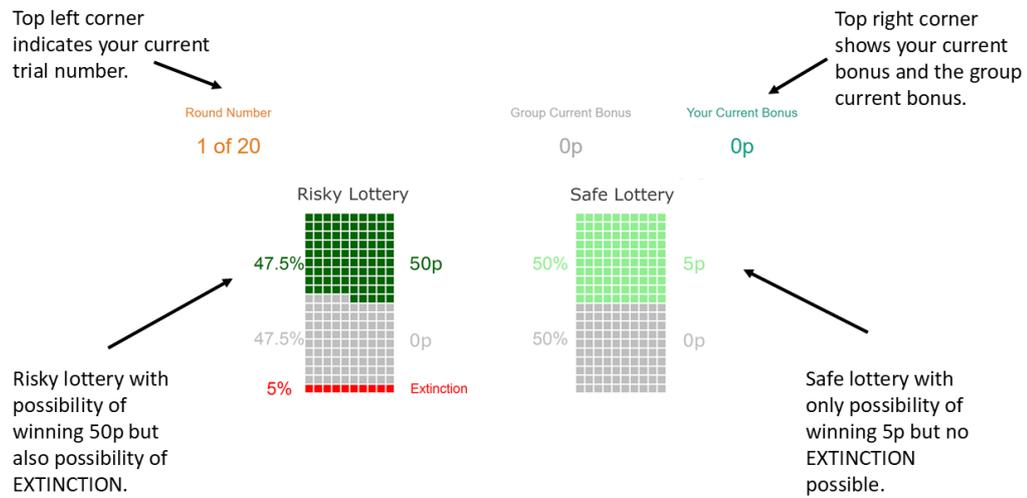
Note. Each player indicated a number between 0 and 5 for how many risky choices the group should play, the median of the ordered votes was then executed and the risky vs. safe choices and associated payoffs were randomly assigned to the players.

After reading all the instructions, participants viewed an annotated game screen (see Figure 5.12), which explains all the relevant task features (the two lotteries, current round and number of rounds, individual and group current bonus) on the screen that they see later when making decisions.

Participants started with 20 practice trials in the individual condition and four in the group condition (we use four in the group as this corresponds to 20 total choices and keeps the optimal strategies the same between practice trials).

Participants in the information conditions received information about the optimal strategy after the practice trials. We only provided this information after the practice trials as the practice trials had a very different optimal strategy, given that the number of rounds is much lower, and we wanted to avoid any spillover from the

Figure 5.12: Annotated Task Setup as Shown to Participants in the Collective Conditions



Note. In the individual condition, the maximum trial number was 100 (instead of 20), the group optimal was not be shown as it was not applicable, and the payoffs were 1p and 10p (instead of 5p and 50p).

practice trial optimal strategy information to the main trial optimal strategy information. The optimal strategy information included a sentence stating how many risky choices the group should make based on the optimal expected value strategy and what that implies for how many risky choices they should make individually. Further, we included a plot indicating the expected bonus payoff as a function of group risky choices. We asked the following comprehension checks to make sure participants understand the optimal strategy manipulation: ‘For the group as a whole, the optimal total number of risky lotteries is . . .’ (correct answer: ‘The group in total plays 8 risky strategies over the entire game’) and ‘If each player wants the group to play the optimal number of risky lotteries, how many risky lotteries should each player play on average?’ (correct answer: ‘Each player plays 1 to 2 risky lotteries over the entire game’). Finally, we also included information about the optimal strategy on each page next to the button, which participants used to indicate their decision. This information was updated dynamically based on the current group endowment and the remaining number of trials.

Analysis Plan

We used a logistic mixed effects model which predicts the number of risky choices in each round based on the round number, round number squared (to account for potential nonlinear trends in risky choices during the experiment and implemented using orthogonal polynomials with the `poly()` function in R), condition, as well as interactions between round number, round number squared, and condition. Because the dependent variable was the number of risky choices a group plays within each round, all analyses were on the group level (rather than on the level of participants nested within groups). To compare the collective conditions to the individual condition, we binned the choices for the individual condition for every five trials, which allowed us to analyse them within the same model (i.e., an individual has 20 bins of 5 trials each, which is equivalent to a group with 20 rounds of 5 trials each).

To test the differences between conditions, we used the *estimated average riskiness* between conditions following the analysis for the individual task in part 1 of this chapter. As all of our hypotheses concern different theoretical questions, we did not adjust for multiple comparisons.

Sampling Plan (Planned Study)

We will power our study to detect a condition difference of 0.7 on the logit scale ($\approx 10\%$ on the probability scale). In terms of effect size conventions, a change of 0.7 on the logit scale corresponds to an odds ratio of ≈ 2 , which is considered a small to medium effect size (typically an odds ratio of 1.5 is considered small, and an odds ratio of 2.5 is considered medium; Maher et al., 2013; J. A. Rosenthal, 1996). Given that our manipulations are relatively strong (i.e., the conditions are quite different in terms of the task setup and materials) and based on results for the individual level task, which usually found effects of 1 or larger on the logit scale, we consider this effect size a reasonable lower bound on the expected effect size if an effect is present. Appendix C.10 indicates details of a simulation-based power analysis for this effect size, which suggests 95% power to detect a condition difference of 0.7

with a sample size of 110 groups per condition (or 110 participants in the individual condition). This would imply a total sample size of 2310. We will collect data until we reach the target sample size in terms of participants and groups that passed all attention and comprehension checks in each condition (i.e., at least 110 participants or groups per condition, for more details see section ‘Materials and Procedure’). Data collection and analysis will not be performed blind to the conditions of the experiments.

Participants (Pilot Study)

To demonstrate the feasibility of our design, we collected 294 participants for a pilot study on 9th of August 2024. Out of the 294 collected participants, 271 passed all attention checks. We further excluded all groups of participants for which at least one participant failed one of the attention checks. This left us with a final sample of 242, comprising 62 participants in the individual condition and 36 groups of 5 participants. 9 groups were in the control condition, 6 groups in the optimal information condition, 11 in the median voting condition, and 10 in the median voting and optimal information condition.¹⁸ The average bonus paid to participants on top of the base rate was 58p, and the median time taken was 14 minutes.

5.2.2 Pilot Results

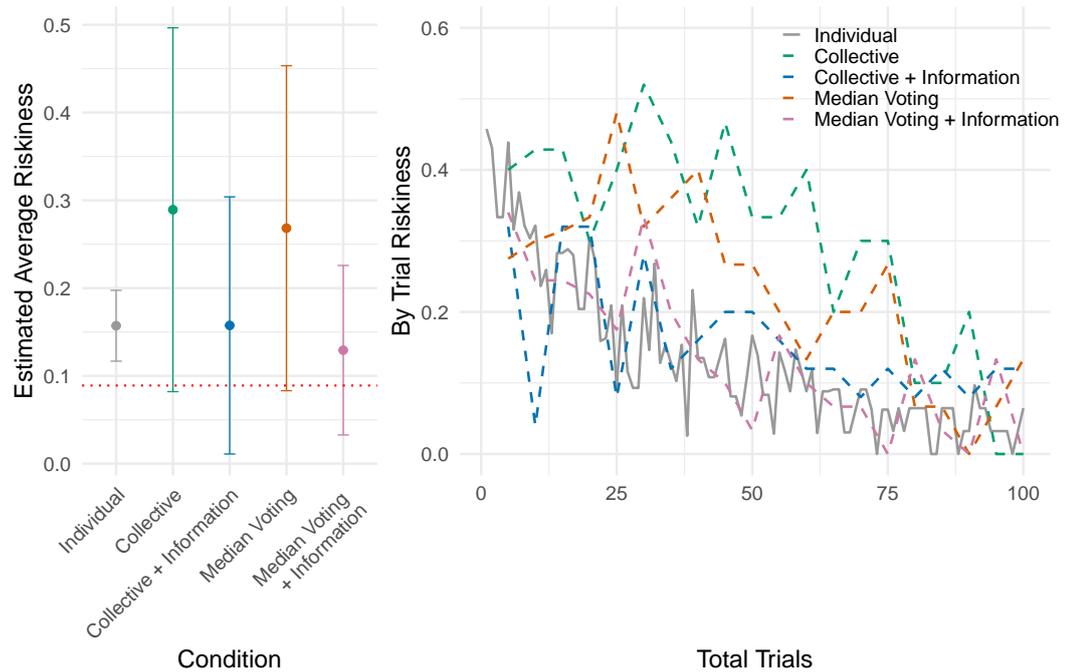
Figure 5.13, Panel A shows the average estimated riskiness between conditions, while Panel B shows the proportion of risky choices over time.

The inferential tests lead to the following conclusions regarding the hypotheses outlined in Table 5.1:

- **H1:** We found a significant difference in the proportion of risky choices between the individual and the collective condition, $\mu_{\text{individual}} = .16$, 95%

¹⁸The unequal assignment is due to the fact that participants were assigned to each group with an equal probability, which creates some sampling variation. In the main study, the relative difference in the number of groups per condition would be much smaller due to the larger total sample size. For the main study, we will also adjust the probability of assignment to the individual condition, to obtain an equal number of individuals as groups for the other conditions (this is desirable based on how we cluster people in our analysis plan, see section ‘Analysis Plan’).

Figure 5.13: Estimated Average Riskiness and Per Trial Proportion of Risky Choices



Note. The dotted red line indicates the optimal level of risky choices.

CI [.12, .20], $\mu_{\text{collective}} = .29$, 95% CI [.08, .50], $z = 1.82$, $p = .035$, one-sided. This suggests that increased risk-taking about extinction risk is partially driven by problems in coordination and cooperation.

- **H2:** We found no significant difference in the proportion of risky choices between the collective + information and the collective condition, $\mu_{\text{collective + information}} = .16$, 95% CI [.01, .30], $\mu_{\text{collective}} = .29$, 95% CI [.08, .50], $z = 1.44$, $p = .075$, one-sided. In other words, there is no evidence that providing information about the optimal strategy reduces risk taking in collective decisions under extinction risk.
- **H3:** We found no significant difference in the proportion of risky choices between the median voting and the collective condition, $\mu_{\text{median voting}} = .27$, 95% CI [.08, .45], $\mu_{\text{collective}} = .29$, 95% CI [.08, .50], $z = 0.25$, $p = .403$, one-sided. In other words, there is no evidence that a median voting intervention reduces risk taking in collective decisions under extinction risk.

- **H4:** We found no significant difference in the proportion of risky choices between the combined condition and the median voting only condition, $\mu_{\text{median voting + information}} = .13$, 95% CI [.03, .23], $\mu_{\text{median voting}} = .27$, 95% CI [.08, .45], $z = 1.46$, $p = .072$, one-sided, or the combined condition and the information only condition, $\mu_{\text{collective + information}} = .16$, 95% CI [.01, .30], $z = 0.03$, $p = .490$, one-sided. In other words, there is no evidence that a combination of the two interventions is more effective than each of them in isolation.

Exploratory Analysis. To increase statistical power for detecting the effect of providing information about the optimal strategy, we conduct an additional exploratory analysis comparing the two collective conditions without information about the optimal strategy (collective and median voting) to the two conditions with information about the optimal strategy (collective + information and median voting + information). This indicates a significant reduction in risk-taking when provided with information about the optimal strategy ($z = 3.39$, $p < .001$).

5.2.3 Conclusion

This section demonstrates how the extinction gambling task can be generalized to study collective decision making. By comparing decision making across an individual-level decision making task and a collective version of the same task with matching optimal strategies, we can understand the extent to which collective versus individual level problems lead to increased risk-taking under extinction risk. More generally, the approach of matching individual and collective tasks introduced here could be generalized to study a variety of societal issues, such as climate change, where both individual-level reasoning limitations and cooperation failures play a role.

While this is only a pilot study for a registered report with a relatively small sample size, it already provides evidence that increased risk-taking under extinction risk is indeed driven by cooperation and coordination problems and exploratory analyses suggest that providing information about the optimal strategy may reduce risk taking. The main study will shed more light on these findings and additional

questions, such as whether median voting can reduce risk-taking in collective decisions. By receiving peer review and in principle acceptance before actually running this study, we can ensure that these findings are robust and free of publication bias.

Corresponding Manuscript: Maier, M.*, Pilgrim C.*, Mann R.P., & Singmann, H. (under review). Collective Decision Making Under Extinction Risk [Stage 1 registered report].

Data and Code Availability: All data and code are available at https://osf.io/cdhfs/files/osfstorage?view_only=1f9b8471ea224394902db9acc0fe7405.

Author contributions: M.M. came up with the idea for the study and derived the optimal strategies for the different games. M.M. and C.P. conceptualised the collective task, and C.P. programmed the experiment in oTree. All authors were involved in the conceptualisation and design of the experiments. M.M. conducted the data analysis and created the visualisations with feedback from H.S. M.M. and C.P. wrote the first draft of the manuscript. All authors contributed to reviewing and editing the manuscript.

Chapter 6

Moral Learning

In the previous chapters, the ethically right decision was relatively straightforward (donate to a charity helping children in need and reduce extinction risk); however, in many real-world decisions, it is not so easy to determine what is morally right, and different people disagree on what the morally best decision would be. In these situations, competing moral theories or decision mechanisms have different implications for which actions someone should choose. In this chapter, I try to understand how people's often conflicting beliefs and intuitions about moral dilemmas arise. To do so, I develop a new learning paradigm and computational models to investigate how people learn to make decisions about dilemmas where moral rules (often equated with deontology) and cost-benefit reasoning (often equated with utilitarianism) clash.

6.1 Metacognitive Moral Learning in Realistic Moral Dilemmas

In the courtroom drama *Terror*, the audience must judge the actions of Major Lars Koch, a fighter pilot accused of killing 164 people. Koch disobeyed orders and shot down a hijacked passenger jet headed towards a stadium filled with 70,000 people. Koch's decision to sacrifice the smaller group to save the larger one was based on utilitarian moral reasoning, but about 36% of the people who saw the play decided that he was guilty (https://terror.theater/cont/results_main/en).

Although such life-or-death decisions are rare in everyday life, people often

face analogous moral dilemmas between following moral rules (e.g., telling a friend the truth about their bad cooking) versus cost-benefit reasoning (CBR; e.g., telling a white lie so as not to hurt the friend's feelings). CBR sometimes endorses violating rules for (what is perceived to be) the greater good. People's decisions in these moral dilemmas have consequences that they can learn from. Moral rules and CBR also clash on a number of important issues, including vaccination mandates and animal testing. These issues often become divisive and highly controversial because people vehemently disagree about whether moral rules take precedence over CBR or vice versa.

The question of whether to rely on moral rules or CBR is often conflated with the normative problem of whether morality consists in choosing actions with good consequences or whether the rightness of an action is inherent in the action itself (Gawronski & Beer, 2017; T. John, 2023; Smart & Williams, 1973). One may therefore believe that maximizing the good consequences of a decision is always achieved through relying on CBR. However, that is not necessarily the case. For instance, in *Terror*, Koch's 'utilitarian' action may have prevented the passengers from stopping the terrorists and saving everyone, and it could also have weakened the crucial general norm against killing. If so, the consequences of following the rule would have been better. As such, although Koch's decision to commit sacrificial harm was based on CBR, it is unclear whether it met the utilitarian criterion to produce the best consequences.

More generally, from a utilitarian perspective, following rules can be viewed as a heuristic that leads to better consequences in certain environments in which CBR is likely to misestimate possible consequences (Bennis et al., 2010; Gigerenzer, 2008). The idea that relying on rules is better from a consequentialist perspective has also been discussed in moral philosophy under rule utilitarianism (E. G. Williams, 2023) or global consequentialism (Ord, 2009). Therefore, in this chapter, we delineate reliance on CBR versus moral rules as *decision strategies* from the endorsement of the ethical theories of deontology and consequentialism. Importantly, we use CBR to refer to a 'naive' cost-benefit reasoning, which considers the num-

ber of persons affected by one or more salient outcome(s) and the corresponding subjective probabilities. We do *not* assume that people engaging in CBR consider all possible consequences, including indirect and long-term consequences, because this kind of exhaustive cost-benefit analysis would be intractable in real-world situations (Gigerenzer, 2008).

What determines how much weight a person puts on moral rules versus CBR in moral dilemmas? One potential mechanism is learning from the consequences of their previous moral decisions. This mechanism is distinct from previous accounts of moral learning (Cushman et al., 2017), including affective learning of moral intuitions (Blair, 2017; Crockett, 2013; Railton, 2017) and moral rules (Nichols, 2021), universalization (Levine et al., 2020), and social learning (Kleiman-Weiner et al., 2017). Unlike social learning, it involves neither imitation nor observational learning and does not require instruction or social feedback (e.g., praise or criticism). Moral learning from consequences is crucial for moral development (Blair, 2017; Gibbs, 2019; P. L. Lockwood et al., 2016, 2024), yet it is comparatively understudied. This chapter therefore makes theoretical and empirical contributions to understanding this type of learning: We develop a formal theory and computational models of an overlooked mechanism of moral learning, provide the first experimental demonstration of its existence and relevance, and introduce an experimental paradigm for studying it.

Our work builds on and extends the reinforcement learning (RL) perspective on moral decision making developed by Cushman (2013) and Crockett (2013). According to this view, CBR and rule-based approaches to moral decision making correspond to two different decision systems: CBR corresponds to the *model-based system*, and rules such as ‘do not kill’ to the *model-free system*.

The model-based system builds a model of the environment and uses it to reason about the potential consequences an action might have in a specific situation. Based on this internal representation, it considers different courses of actions, predicts their outcomes, and plans the courses of actions that produce the most optimal expected outcomes overall. Conversely, rather than relying on a model of the en-

vironment, the model-free system assigns value directly to each action based on its history of reward and punishment. The action's value representation is adjusted through prediction error learning by comparing the previous representation of the value with the current outcome.

Depending on the situation, people's decisions seem to follow one system or the other (Gawronski et al., 2017b). Previous work linking the model-based system to CBR and the model-free system to following rules was concerned with the question of which system determines the actions implied by the strategies of following CBR or rules. However, it does not answer the question of how people choose to follow a certain decision strategy in the first instance; this is a separate question related to strategy selection (Lieder & Griffiths, 2017).

Theories of strategy selection (Erev & Barron, 2005; Lieder & Griffiths, 2017; Rieskamp & Otto, 2006) and meta-control (Boureau et al., 2015; Gershman et al., 2015; Lieder, Shenhav, et al., 2018; Shenhav et al., 2017) postulate an overarching *meta-control* system that decides which decision mechanism to use in a given situation. Based on these theories, we propose that the meta-control system selects which *moral* decision mechanism to employ in a specific situation. Given the strong empirical evidence for the pervasive influence of reinforcement learning (RL) on decision making (D. Lee et al., 2012; O'Doherty et al., 2015) and strategy selection (Erev & Barron, 2005; Lieder & Griffiths, 2017; Lieder, Shenhav, et al., 2018; Rieskamp & Otto, 2006; Verbeke & Verguts, 2024), we postulate that meta-control over moral decision making is also shaped by RL (*metacognitive moral learning*).

In the remainder of this chapter, we formalize this hypothesis and test it in two experiments. In Experiment 1, we demonstrate the existence of adaptive metacognitive moral learning from the consequences of previous decisions, show that it transfers to real-life, incentive-compatible donation decisions, and find that metacognitive learning is a requirement for this transfer. In Experiment 2, we rule out that the findings are due to demand characteristics by demonstrating transfer for metacognitive learners to a different experiment, which participants thought was conducted by different researchers.

6.1.1 A Theory of Metacognitive Moral Learning

Prior work has identified several mechanisms of moral learning (Cushman et al., 2017). According to one of these mechanisms, reinforcement learning (RL), moral values are learned from the consequences of previous actions. Each time the consequences of an action are better than expected, the probability of repeating this action is increased, and each time the action's consequences are worse than expected, the probability of repeating this action is reduced. While prior theories of moral learning (Crockett, 2013; Cushman, 2013) proposed that people learn on the level of more specific behaviours (e.g., whether to punch someone), we propose that people also learn on the level of moral decision-making strategies (e.g., whether to engage in rules or CBR). We refer to this mechanism as *metacognitive moral learning*.

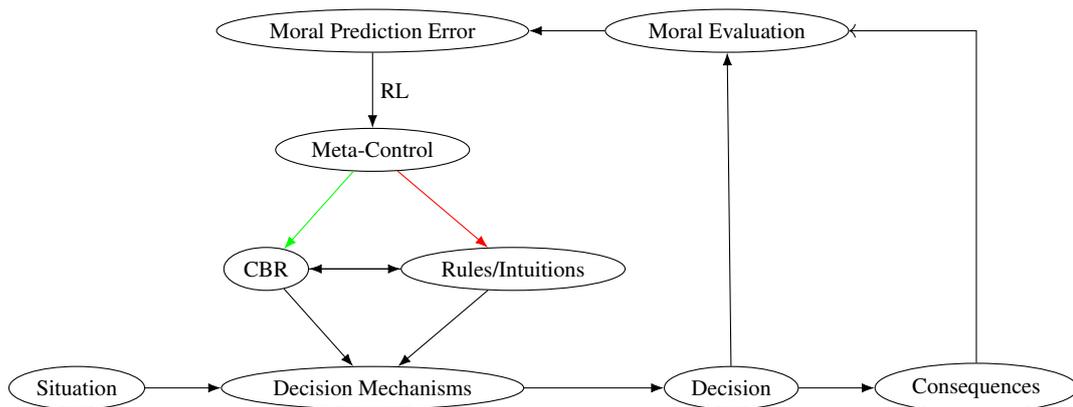
According to this theory, the mechanisms of strategy selection learning (Erev & Barron, 2005; Lieder & Griffiths, 2017; Lieder, Shenhav, et al., 2018; Rieskamp & Otto, 2006) also operate on the mechanisms of moral decision making. In strategy selection learning, the consequences of people's actions reinforce the decision strategies that selected them, unlike in operant conditioning (Skinner, 1963), where consequences reinforce specific behaviours. Therefore, we propose that when a person concludes that one of their past decisions was morally wrong (right), this will teach them to decrease (increase) their reliance on the decision system or strategy that chose that action (e.g., rule-following or CBR; see Figure 6.1). For example, in *Terror*, the audience learns not only about the morality of shooting down airplanes, but also about the morality of CBR more generally. Importantly, if people only learned about specific behaviours, moral learning would not generalise to different *types* of behaviours. By contrast, *metacognitive* moral learning should transfer to novel situations involving other behaviours.

Put simply, the mechanisms of metacognitive moral learning differ from standard RL in two key ways: (1) learning occurs in the meta-control system whose 'actions' are our decision strategies (e.g., CBR and rule-following), and (2) the reward signal is the person's moral evaluation of how good or bad their decision was. These moral evaluations are partly based on consequences of the decision (Gawron-

ski et al., 2017b, 2023). Therefore, learning from the consequences of past decisions could, in principle, adaptively increase people’s reliance on decision strategies that produce good consequences, and decrease reliance on those that produce bad consequences.

Given the coexistence of model-free and model-based RL (Dolan & Dayan, 2013), we postulate that metacognitive moral learning includes both model-based and model-free RL mechanisms. Model-free metacognitive moral learning consists of learning the *expected moral values* of relying on different decision strategies. By contrast, model-based metacognitive moral learning consists of learning *conditional probability distributions* over the possible *outcomes* of relying on different decision strategies (see Section 6.1.3).

Figure 6.1: Meta-Control of Moral Decision-Making is Informed by Learning from Previous Decisions.



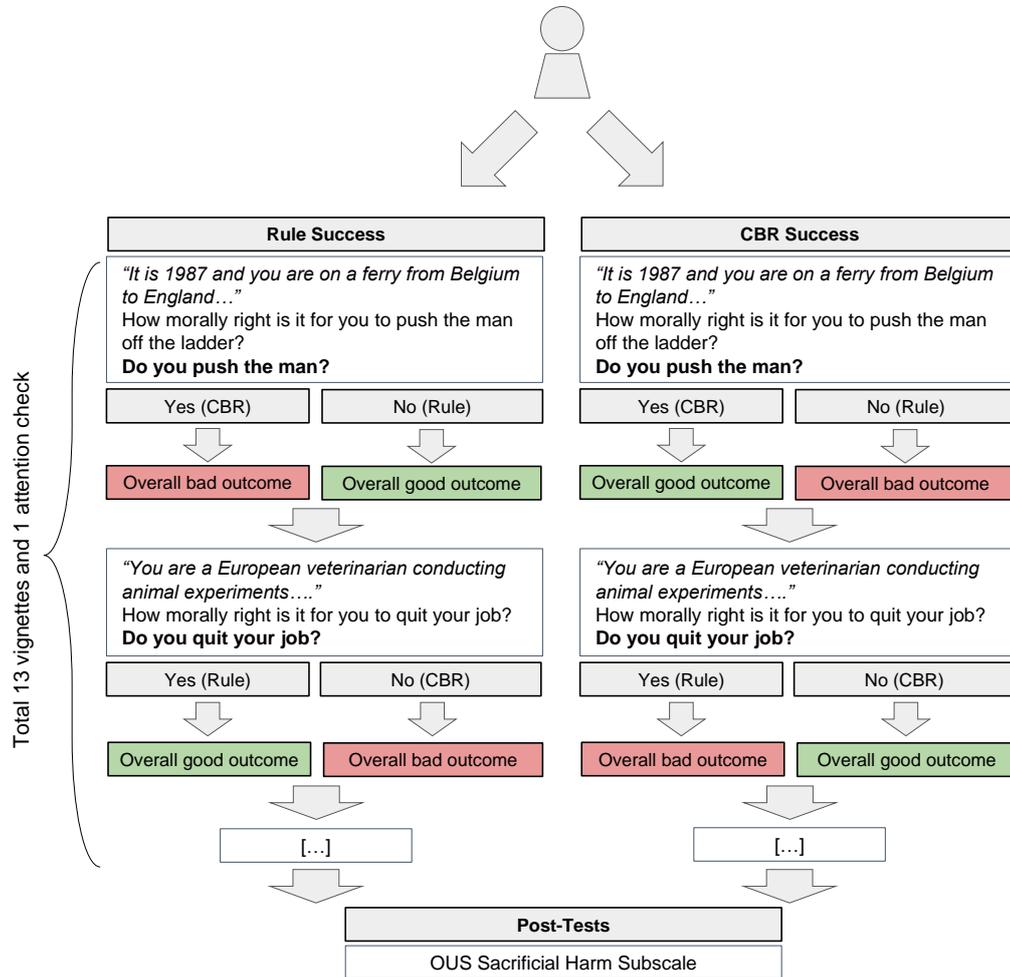
Note. The meta-control system determines which of multiple decision mechanisms, including moral rules and cost-benefit reasoning (CBR), is employed in a particular situation (in this chapter, we focus on rules vs. CBR, though the model could be extended to accommodate other mechanisms). Depending on one’s learning history, the meta-control system may temporarily override moral rules (red arrow) by allocating control over behaviour to CBR (green arrow) or vice versa. Whether moral learning increases (decreases) reliance on CBR or rules in subsequent decisions depends on how positively (negatively) one evaluates the previous decision. People’s moral evaluation of past decisions are influenced by the consequences of these decisions. If the evaluation is more (less) positive than expected, this registers as a positive (negative) moral prediction error that causes the reliance on the mechanism that produced the decision to be turned up (down). We suggest that this strategy selection learning shapes moral decision making.

6.1.2 A New Experimental Paradigm Using Realistic Trolley-Type Dilemmas with Outcomes

To test our theory and models of metacognitive moral learning, we developed an experimental paradigm for measuring the effect of learning from the consequences of previous moral decisions on subsequent moral decisions. Unlike previous moral decision making paradigms, ours is a learning paradigm. Participants make decisions in a series of different moral dilemmas, where they see the outcomes of each decision before moving on to the next. In each trial, the participant reads a realistic moral dilemma and decides between two actions: one favored by CBR (the ‘CBR option’) and one favored by a moral rule (the ‘rule option’).

At the beginning of the paradigm, participants are randomly assigned to one of two conditions. In the ‘CBR Success’ condition, the CBR option always leads to overall good outcomes, and the rule option to overall bad outcomes. In the ‘Rule Success’ condition, the rule option always leads to overall good outcomes, and the CBR option to overall bad outcomes. We illustrate this paradigm in Figure 6.2, and more details can be found in the Method section.

Figure 6.2: The Moral Learning Paradigm in Experiment 1.



Note. Participants are randomized into one of two conditions: 'Rule Success' and 'CBR Success'. The experimental condition determines whether choosing the CBR option or the rule option leads to good or bad outcomes. In some vignettes, the action under consideration is the CBR option (e.g., 'Do you push the man?'), and in other vignettes, it is the rule option (e.g., 'Do you quit your job?'). This means that the 'yes' and 'no' responses do not always correspond to the same decision strategy (rules vs. CBR) even in the same experimental condition. Because of this, from the participant's perspective, choosing 'yes' would sometimes lead to good outcomes and sometimes lead to bad outcomes. More details about the paradigm can be found in the Methods section.

The moral dilemmas most widely used in experiments, which are based on the 'trolley problem' (Foot, 1967; Thomson, 1976), have been criticized as unrealistic and bizarre (Bauman et al., 2014; Bennis et al., 2010). Further, they assume that the outcomes are known with certainty and often confound CBR with taking action and rule-following with inaction (i.e., omission). The moral dilemmas used in our

paradigm mitigate all of these limitations.

Most participants are not trained in moral philosophy, meaning that the abstract moral theories of deontology and utilitarianism are likely less salient for them than the concrete choices between action versus omission, the specific behaviours, and the specific moral rules that recommend or oppose them. Our experimental paradigm varied all of these salient features independently of which option each strategy recommended (see Figure 6.2 and the Methods section for more details). Therefore, it is not immediately obvious to participants what is being reinforced in our paradigm. This was also reflected in their responses to an open question in a previous experiment.¹

Further, a majority of participants engaged strongly with the task and considered it informative about the real world: 90% of participants reported that they imagined the scenarios very vividly and 90% reported that they felt good or bad after they saw good or bad outcomes. 67% of participants indicated that the decisions, situations, and outcomes they encountered in the task were informative about the real world, and 50% of participants indicated that the task gave them the opportunity to learn how to make better decisions in the real world. Finally, most participants indicated that the outcomes were plausible (83%) and a good reflection of whether they made the right decision (61%; see supplementary information for more details). For more details on the paradigm see also the Method section of (Maier, Cheung, & Lieder, 2024).

¹At the end of Experiment 1 of the corresponding journal article, we asked participants whether they ‘used information about the outcomes of [their] choices when making decisions throughout the experiment,’ and, if so, how. Of those participants who reported taking outcomes into account, most appeared to be unaware of the specific manipulation (e.g., ‘Yes I tried to worry more about the initial moral decision and less on the outcomes as it was clear the outcome could vary/was more unpredictable’ and ‘I tried to anticipate what the likely outcome would be, but I wasn’t right’; full responses are available in the online repository).

6.1.3 Computational Models of Moral Learning from Consequences

General Background

To test our theory, we developed RL models of metacognitive moral learning from the consequences of past decisions. As metacognitive learning could be model-based or model-free, we developed one computational model to represent each.

Model-based learning uses an explicit model of the world to estimate the conditional probabilities of different outcomes (Doll et al., 2012, Sutton and Barto, 2018, p. 159). We modeled model-based metacognitive moral learning as Bayesian learning of the conditional probabilities of good versus bad outcomes of decisions made by CBR versus following moral rules (e.g., $P(\text{good outcome} \mid \text{CBR})$). This model learns these probabilities by updating the parameters of two beta distributions: one for the probability that CBR will yield a good outcome and one for the probability that following rules will yield a good outcome. The probability of a bad outcome is simply one minus the probability of a good outcome. In other words, this model estimates the likelihoods of four different outcomes: following CBR leads to good versus bad outcomes, and following rules leads to good versus bad outcomes.

Model-free learning assigns values to actions directly, rather than modelling the probability of different outcomes. Those values are based on the average reward each action produced in the past. To model model-free metacognitive moral learning, we adapted the most common model of model-free RL: Q -learning (Dearden et al., 1998; Watkins, 1989; Watkins & Dayan, 1992). Our model assigns values directly to using moral decision-making strategies (i.e., CBR vs. following moral rules); those values are represented as Q -values. After each decision, the model updates the Q -value of relying on the decision strategy that produced that decision. This update is proportional to the experienced moral prediction error. The moral prediction error is the difference between the decision-maker's moral evaluation of how morally right or wrong the decision was and the current Q -value of the decision strategy that produced it. The higher the Q -value assigned to a decision strategy, the

more likely the model is to rely on it.

Unlike the model-based Beta-Bernoulli model, the model-free Q -learning model does not learn about the probabilities of the four different outcomes, but instead learns two Q -values: one for CBR and one for rule following. Therefore, our two computational models capture the key distinction between model-based and model-free learning: model-based learning involves learning about the probabilities of the different outcomes of an action, whereas model-free learning assigns a value to the action itself.

Our models of metacognitive moral learning attribute the outcome of each action to the decision strategy that selected it (i.e., applying CBR vs. moral rules). We compared these models to models of behavioural moral learning. Unlike metacognitive learning, behavioural learning attributes the outcome of each action to the action itself. For example, a child pushing their friend out of the sandbox may see that this action causes their friend to become upset, and learn not to repeat such actions. To model the generalisation of behavioural learning across the different dilemmas of our experimental paradigm, we make the simplifying assumption that people generalise from the outcome of (not) taking the action under consideration in any one dilemma to the value of (not) taking the action under consideration in all other dilemmas. Our models of behavioural learning thus assume that each decision is represented as either performing the behaviour under consideration (action) or not (omission). Actions were a very salient behaviour level representation on which the learning signal could operate, given that in each trial, participants were asked whether to act (e.g., push the man) or not.

Behavioral learning can be either model-based or model-free. Apart from changing the learning signal to operate on the level of behaviours rather than strategies, our models of model-based versus model-free behavioural learning are therefore equivalent to our models of model-based versus model-free metacognitive learning. Our paradigm allows deconfounding action/omission learning from metacognitive learning, as sometimes action coincides with CBR and sometimes with rules.

Model Specification

Model-Free Metacognitive Learning (Q-Learning). We formalize model-free metacognitive learning by a Q -learning model of how people solve the meta-control problem of deciding when to rely on moral rules versus CBR (see Figure 6.1). This model learns to predict the anticipated moral value $Q^{\text{meta}}(s, \text{CBR})$ of relying on CBR in the current situation s and the moral value $Q^{\text{meta}}(s, \text{rules})$ of relying on moral rules. Assuming that in trial t control was allocated to CBR then, once the consequences of the resulting action are observed, the model calculates the moral prediction error,

$$\text{MPE}_t = Q_t^{\text{meta}}(s, \text{CBR}) - \text{MJ}_t, \quad (6.1)$$

which is the difference between the model's prediction of the moral value of relying on CBR ($Q_t^{\text{meta}}(s, \text{CBR})$) and the person's moral evaluation of how good their decision was after they observed its consequences (MJ_t). The participant provided a rating on a scale from -100 to +100. To obtain MJ_t , we divided that rating by 100.

The model then uses the moral prediction error (MPE) to update its estimate of the moral value of using CBR according to Equation 6.2. How strongly this prediction is updated depends on the learning rate (α).

$$Q_{t+1}^{\text{meta}}(s, \text{CBR}) = Q_t^{\text{meta}}(s, \text{CBR}) - \alpha \times \text{MPE}_t, \quad (6.2)$$

Conversely, when the decision was made by applying a moral rule, then the equivalent update was applied to the estimated value of rule-following (i.e., $Q_t^{\text{meta}}(s, \text{rules})$).

In each new decision situation, the learned values of relying on CBR or rules determine the probability that the decision-maker will engage in CBR or apply moral rules according to the softmax decision rule specified in Equation 6.3.

$$p_t(s, \text{CBR}) = \frac{e^{\tau \times Q_t^{\text{meta}}(s, \text{CBR})}}{e^{\tau \times Q_t^{\text{meta}}(s, \text{CBR})} + e^{\tau \times Q_t^{\text{meta}}(s, \text{rules})}} \quad (6.3)$$

The parameter τ (inverse decision temperature) controls how deterministically the

meta-controller allocates control to the decision mechanisms that it expects to produce morally better outcomes. Larger values of τ imply more deterministic meta-control, whereas lower values of τ imply more random meta-control.

As the prior distribution on the temperature parameter τ , we use $\text{lognormal}(0, 1.4)$. We chose this prior distribution because it assigns 90% of the prior probability mass to values of τ between $\frac{1}{10}$ and 10. The prior distribution on the learning rate is a uniform distribution on the interval $[0, 1]$. This prior reflects the belief that learning rates larger than 1 (i.e., changing your belief by more than the prediction error) and learning rates smaller than 0 (i.e., learning the opposite of what the prediction error suggests) are impossible.

Model-Based Metacognitive Learning (Beta-Bernoulli Updating). For the model-based learning model, we assume that the meta-controller learns the probabilities that a decision made by CBR versus rules will lead to a good versus bad state (e.g., $P(s' \in \mathcal{G} | s, \text{CBR})$, where \mathcal{G} is the set of good states). We model this as Bayesian learning with conjugate priors. Concretely, for each decision mechanism, the prior is a beta distribution on the probability θ_t^{CBR} that the outcomes will be good overall. The likelihood function is a Bernoulli distribution over two possible outcomes: $+1$, meaning that the outcome was good overall, and -1 , meaning the outcome was bad overall.

Thus, after learning from the person's moral evaluation $\text{MJ}_1, \dots, \text{MJ}_t$ of the decisions made using the selected decision mechanisms M_1, \dots, M_t in trials $1, \dots, t$, the model's posterior distribution ($P(\theta_t^{\text{CBR}} | \text{MJ}_1, \dots, \text{MJ}_t, M_1, \dots, M_t)$) on the probability of good outcomes resulting from CBR is

$$\text{Beta} \left(\alpha + \sum_{i=1}^t \mathbb{1}(C_i = \text{CBR and } R_i > 0), \alpha + \sum_{i=1}^t \mathbb{1}(C_i = \text{CBR and } R_i < 0) \right), \quad (6.4)$$

where $\mathbb{1}(\cdot)$ is the indicator function, which is one if and only if its argument is a true statement, and α determines the strength of the prior belief that positive and negative outcomes are equally likely. For higher values of α learning is slower. α , therefore, serves a similar function as the learning rate of the Q learning model.

Conversely, after the first t trials, the model's posterior distribution ($P(\theta_t^{\text{rules}} \mid \text{MJ}_1, \dots, \text{MJ}_t, M_1, \dots, M_t)$) over the probability that relying on rules will produce a good outcome is

$$\text{Beta} \left(\alpha + \sum_{i=1}^t \mathbb{1}(C_i = \text{rules and } R_i > 0), \alpha + \sum_{i=1}^t \mathbb{1}(C_i = \text{rules and } R_i < 0) \right). \quad (6.5)$$

The decision mechanism is again selected using a softmax decision rule based on the learned posterior probabilities of each decision mechanism producing morally good versus morally bad outcomes, as specified in Equation 6.6.

$$p_t(s, \text{CBR}) = \frac{e^{\tau \times \theta_i^{\text{CBR}}}}{e^{\tau \times \theta_i^{\text{CBR}}} + e^{\tau \times \theta_i^{\text{rules}}}}. \quad (6.6)$$

For the prior distribution on τ , we again use $\text{lognormal}(0, 1.4)$. For the prior distribution on the prior precision α , we use $\text{Gamma}(\text{shape} = 2.57, \text{rate} = 0.54)$. This distribution assigns 90% of the probability mass to values between 1 and 10, which is equivalent to having seen between 1 and 10 instances of a positive outcome and between 1 and 10 instances of a negative outcome before starting the experiment.

Models of Model-Free and Model-Based *Behavioral* Learning. Our models of model-free and model-based *behavioural* learning were exactly analogous to the models of model-free and model-based *metacognitive* learning described above. The only difference was that the learning rules that the metacognitive models use to learn the value or transition probabilities associated with relying on alternative decision mechanisms (CBR vs. following moral rules) are applied to the value or transition probabilities associated with the person's behaviour (action vs. omission).

Our model of model-based behavioural learning estimates one single state-transition probability for all behaviours that the vignettes framed as the action under consideration and one single state-transition probability for not performing those behaviours. That is, model-based behavioural learning computes two posterior distributions: one for the probability that taking the action under consideration will lead to a good state (i.e., θ_t^{action}) and one for the probability that not taking that

action will lead to a good state (i.e., $\theta_t^{\text{omission}}$).

Constant Probability Models: No Learning. As a baseline for our models of learning, we formulated equivalent models of what decisions people would make if there was no learning. The baseline for models of metacognitive learning assumes that the probability of relying on CBR versus rules is constant over time, that is

$$p_t(s, \text{CBR}) = \theta_{\text{CBR}}, \quad (6.7)$$

where θ_{CBR} is a free parameter with a uniform prior, that is

$$\theta_{\text{CBR}} \sim \text{Uniform}([0, 1]). \quad (6.8)$$

The baseline for models of behavioural learning assumes that the probability of performing the behaviour under consideration is constant over time, that is

$$p_t(s, \text{action}) = \theta_{\text{action}}, \quad (6.9)$$

where θ_{action} is a free parameter with a uniform prior, that is

$$\theta_{\text{action}} \sim \text{Uniform}([0, 1]). \quad (6.10)$$

Model Implementation

We implemented all models using `Stan` and `RStan` (Carpenter et al., 2017; Stan Development Team, 2024), fitted them separately for each participant, and evaluated the marginal likelihoods using the `bridgesampling` R package (Gronau et al., 2020a). We then used `bmsR` (Elisi, 2024) to conduct Bayesian model selection. Since we developed multiple models of metacognitive learning, multiple models of behavioural learning, and multiple models of moral decision making without learning, we compared the proportions of participants best explained by each model family using family-level inference using the MATLAB function `spm_compare_families` (Penny et al., 2010; Rigoux et al., 2014) with 100,000

samples. We used the same approach to compare the proportions of people best explained by model-based versus model-free learning.

Model Recovery Simulations

We verified that all models could in principle be recovered by simulating from each of the six models, fitting all models on the simulated dataset, and checking whether the marginal likelihood of that model, which we simulated from was indeed the largest. This indicated that the model comparison consistently recovered the model that was being simulated from. We share the code in the online repository.

6.1.4 Experiment 1

In Experiment 1, we tested whether people learn to adjust their moral decisions based on the consequences of their previous decisions and to what extent this learning is metacognitive (i.e., operating on decision mechanisms, such as rules versus CBR, rather than more specific behaviours). Further, we tested how different types of learning transfer to measures of moral convictions and donation decisions.

Method

This experiment received ethical approval from the Office of the Human Research Protection Program, The University of California, Los Angeles (UCLA OHRPP) under protocol number IRB#23-001436. The experiment and data analysis were preregistered at <https://osf.io/7guj6> on the 13th of January 2024.

Participants. We recruited 900 UK-based participants from Prolific on the 13th of January 2024 based on an a priori power analysis, which indicated sufficient power with 429 participants per condition for $d = 0.17$ on the transfer measures (based on a previous experiment reported in Maier, Cheung, and Lieder, 2024). Participants were pre-screened for fluency in English and their past approval rate ($\geq 95\%$) and experience (having previously participated in at least ten studies) on the platform. We paid participants £4.76 (US\$6) for the 36-minute study (base rate of \$5, and to encourage careful reading, a bonus payment of \$1 for passing the attention check).

As preregistered, we excluded participants who failed the attention check ($N = 66$, the attention check was a fake moral dilemma in which participants were instructed to take a clearly inferior action). Our final sample size was $N = 834$ ($M_{\text{age}} = 43.4$, $SD_{\text{age}} = 14.2$, $N_{\text{female}} = 424$, $N_{\text{male}} = 410$).

Materials. After reading through the instructions (and passing a comprehension test), participants completed 13 trials of the moral learning paradigm described in the section ‘A New Experimental Paradigm Using Realistic Moral Dilemmas’ (CBR Success: $N = 420$; Rule Success: $N = 414$). After the paradigm, participants completed self-report measures of moral convictions (OUS Sacrificial Harm Subscale, Kahane et al., 2015, and a four-item questionnaire about deontological decision making taken from the Deontological-Consequentialist Scale, Mata et al., 2022) and a donation task in randomized order.

For the donation task, participants made a series of three donation decisions. Each time, they were asked to allocate £200 between two charities, one of which would be endorsed by cost-benefit reasoning and one of which would be endorsed by rule-based decision making (for more details on how we piloted the charities, see supplementary materials). The three pairs of charities used in the three allocation decisions were:

- ‘Human Challenge Trials’: Choice between 1Day Sooner, a charity promoting human challenge trials in which healthy volunteers are infected with a virus to speed up the development of vaccines (CBR option) vs. the Medical Research Foundation, a charity supporting conventional medical research (rule option).
- ‘Animal Testing’: Choice between Breast Cancer Now, an organization funding animal research to combat breast cancer (CBR option) vs. Breast Cancer UK, an organization that does not fund animal research (rule option).
- ‘Doctors’: Choice between UK-Med, a charity providing humanitarian and medical aid by deploying healthcare professionals to conflict or disaster zones that are inherently risky (CBR option) vs. Pathway, a charity that focuses on

healthcare for homeless people in the UK (rule option).

Participants made these allocations knowing that the experimenters would execute the allocation decision of one randomly chosen participant for one randomly chosen pair of charities.² The full vignettes are available on the supplementary materials.

At the end of the study, as exploratory measures, we included the Empathic Concern and Emotional Empathy scales (Jordan et al., 2016). We also included six items from a Matrix Reasoning task (Condon & Revelle, 2014) to measure cognitive ability.

Data Analysis. We followed the preregistered data analysis plan unless specifically noted. To test the effect of trial number on choices, we preregistered a logistic mixed effects model predicting the probability of the utilitarian choice from trial number and action framing.

For analysing the moral judgments, we preregistered to test the interaction of trial number and framing within a linear mixed effects model that also included the appropriateness ratings from Körner and Deutsch (2023) as a covariate.

To test overall donation behaviour, we test the main effect of donations to CBR charities between conditions. As preregistered, we use a model with random intercepts, main effect of condition and donation type, and an interaction between condition and donation.

To test moderation effects and the effects for those participants who show evidence for metacognitive learning, we first calculate the evidence for metacognitive learning for each participant. We do this by calculating an inclusion Bayes factor comparing the posterior odds of models that describe strategy learning (MF-M and MB-M) with models that describe behavioural learning from action/omission (MF-B and MB-B) or no learning (C-M and C-B). The code for calculating the inclusion Bayes factors is available in the online repository and the preregistration. We estimated marginal likelihoods using the `bridgesampling` package (Gronau et al.,

²We later executed one participant's decision to donate £150 to 1Day Sooner and £50 to Medical Research Foundation on February 16, 2024.

2020b). For more information on inclusion Bayes factors, see Hinne et al. (2020), Maier, Bartoš, and Wagenmakers (2023), and Maier, Bartoš, et al. (2024).

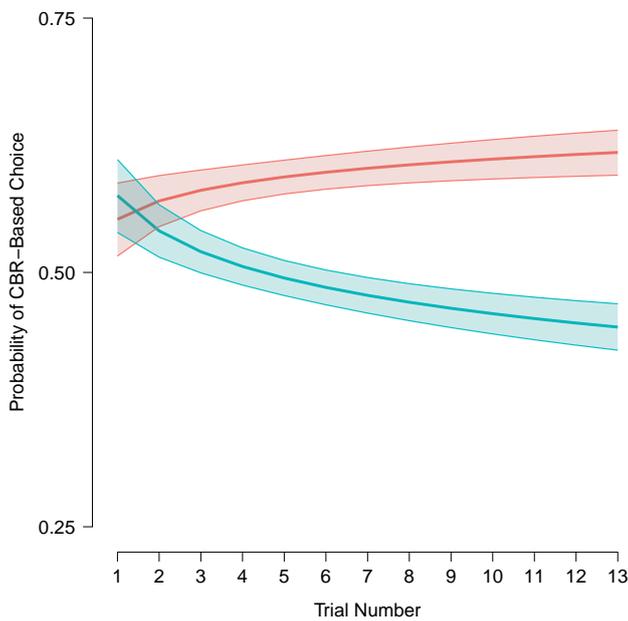
To identify transfer effects only for metacognitive learners, we test the effect of condition in a model that uses evidence for metacognitive learning (in terms of an inclusion Bayes factor contrasting MB-M and MF-M with the other four models) as a covariate. We then recenter the inclusion Bayes factor covariate so that 0 indicates strong evidence for metacognitive learning. This allows us to test the effect of the experimental condition for those participants who showed evidence for metacognitive learning.

Results

Influence of Outcomes on Choices and Judgments. Participants learned from the consequences of previous moral decisions (Figure 6.3): They became more reliant on CBR in the CBR Success condition ($b = 0.11, \chi^2(1) = 7.67, p = .003$) and more reliant on moral rules in the Rule Success condition ($b = 0.20, \chi^2(1) = 28.34, p < .001$). In terms of moral judgements, we found an effect on the judgements in the Rule Success condition ($b = 1.45, F(1, 2201.94) = 7.61, p = .006$), but not in the CBR Success condition ($b = 0.11, F(1, 4811.43) = .01, p = .930$, we address potential explanations for the inconsistent effect on the ratings in the General Discussion).

Computational Modelling Results. We found that in the CBR Success condition, most participants (86.97%) relied primarily on model-based metacognitive learning. In the Rule Success condition, most participants (56.17%) relied primarily on model-based behavioural learning and only 31.81% engaged in model-based metacognitive moral learning (see Table 6.1). When comparing families of models, in the CBR Success condition, the proportion of participants whose behaviour was best explained by either of the models of metacognitive learning (92.5%) was significantly larger than the proportions of participants best explained by models of behavioural learning or no learning (see Table 6.2). By contrast, in the Rule Success condition, the two models of behavioural learning jointly provided the best explanation for the majority of participants (60.8%), while the two models of metacog-

Figure 6.3: Learning from Consequences Shapes Reliance on Moral Rules Versus Cost-Benefit Reasoning.



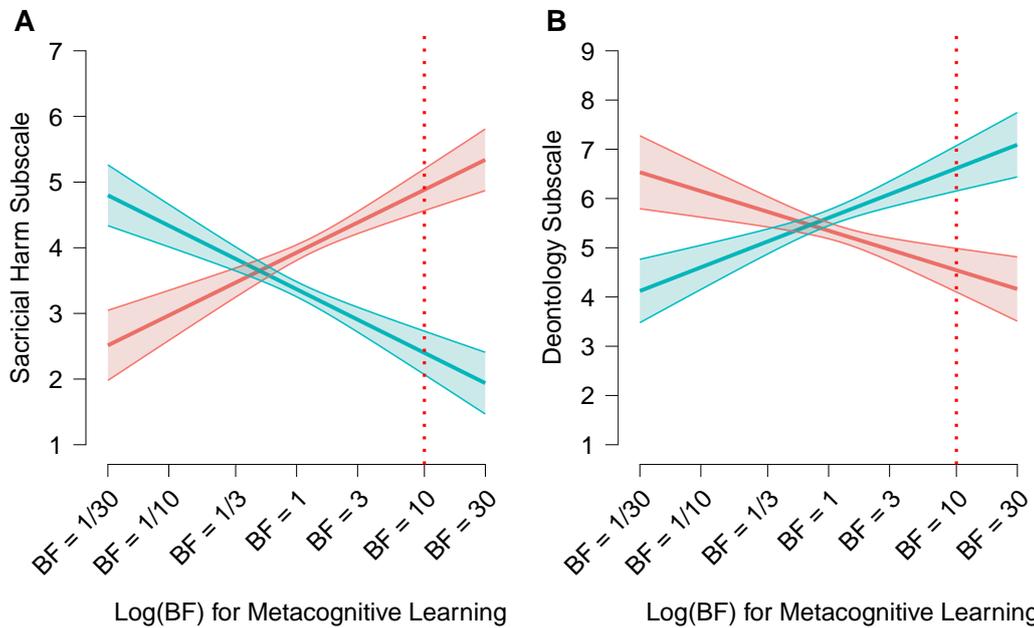
Note. Confidence bands indicate 95% CIs.

nitive learning provided the best explanation for only 35.8% of the participants in that condition (see Table 6.2).

Metacognitive Learning Transfers to Donations and Measures of Moral Convictions. As predicted, evidence for metacognitive learning moderated the effect of the experimental condition on all measures of transfer (OUS Sacrificial Harm Subscale, $b = 0.84, t(830) = 8.39, p < .001$; DCS Deontology Subscale, $b = -0.79, t(830) = 5.67, p < .001$; Donations, $b = 9.15, t(830) = 5.20, p < .001$) and we found strong evidence of transfer for metacognitive learners on all measures of transfer (OUS Sacrificial Harm Subscale, $b = 2.48, t(830) = 10.62, p < .001$; DCS Deontology Scale, $b = -2.07, t(830) = 6.36, p < .001$; Donations, $b = 22.07, t(830) = 5.34, p < .001$, all one-sided).

In contrast, when averaging across all participants (i.e., including those that show evidence for behavioural learning), we found evidence for transfer to people’s moral convictions (OUS Sacrificial Harm Subscale, $t(829.06) = 7.50, p < .001, d = 0.52$, one-sided; DCS Deontology Subscale, $t(825.3) = 2.81, p = .003, d = 0.20$, one-sided) but were unable to detect the effect on donations ($t(832) = 1.55, p =$

Figure 6.4: Endorsement of Sacrificial Harm and Deontology Are Moderated By Evidence for Metacognitive Learning in Experiment 1.



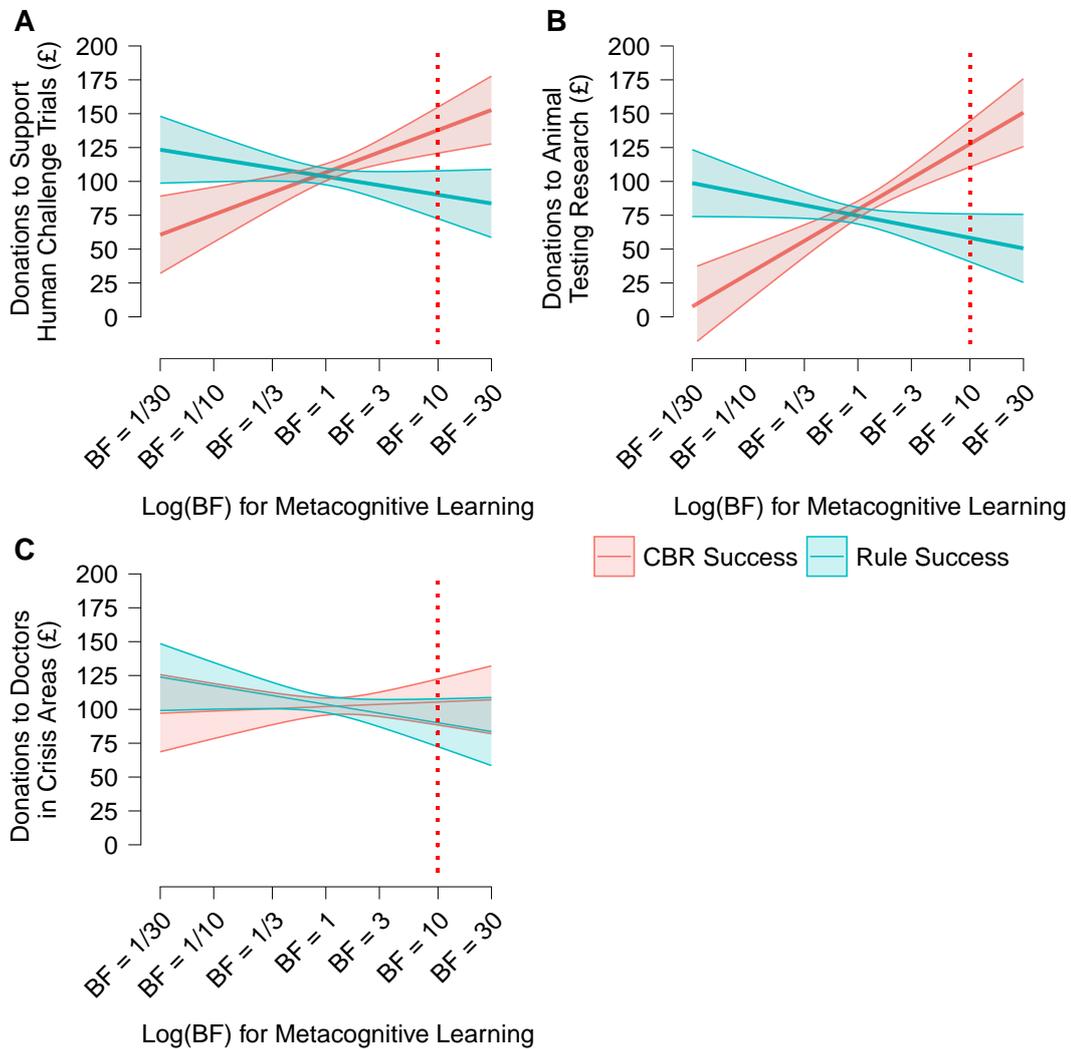
Note. Panel A shows responses to the Sacrificial Harm Subscale from the Oxford Utilitarianism Scale. Panel B shows responses to the Deontology Subscale from the Deontological-Consequentialist Scale. Confidence bands indicate 95% CIs.

.061, one-sided). This discrepancy underscores the importance of metacognitive learning for transfer (see the supplementary information for additional results, including effects on individual donation decisions).

6.1.5 Experiment 2

Experiment 1 showed that moral decision making is driven by learning from consequences and that metacognitive moral learning in our paradigm transfers to other measures within the context of the same experiment. In principle, this could be due to demand characteristics or very narrow, highly context-specific learning. Therefore, Experiment 2 aimed to demonstrate that the effects of moral learning from consequences transfer beyond the experiment in which the learning took place (i.e., transfer to another study conducted by different experimenters). To achieve this, we used an innovative experimental design comprising two separate online studies run by different experimenters from different institutions. The first online study employed the learning paradigm, and the subsequent, seemingly unrelated study

Figure 6.5: Donations to CBR Charities Are Moderated By Evidence for Metacognitive Learning in Experiment 1.



Note. Panel A shows donations to 1Day Sooner, an organization advocating for human challenge trials to accelerate medical research (vs. the Medical Research Foundation, which funds more conventional medical research). Panel B shows donations to Breast Cancer Now, an organization that funds animal testing in research to combat breast cancer (vs. Breast Cancer UK, an organization that does not fund animal testing). Panel C shows donations to UK-Med, a charity that provides medical aid to conflict or disaster zones that have inherent risks (vs. Pathway, a charity that focuses on healthcare for homeless people in the UK, which is comparatively much less risky). More information about each charity is provided in the Methods section. Confidence bands indicate 95% CIs.

measured people's moral convictions and donation behaviour. This allowed us to show that the learning transfers to a new experimental context, thus ruling out the alternative explanation that effects are driven by demand characteristics.

Method

This experiment received ethical approval from the Office of the Human Research Protection Program, The University of California, Los Angeles (UCLA OHRPP) under protocol number IRB#23-001436, and by the UCL Psychology Ethics Committee under code EP/2018/005. The experiment and data analysis were preregistered at <https://osf.io/dgsfb> on 19th of July 2024. Because our theoretical framework predicts transfer only for metacognitive learners and focusing on metacognitive learners had higher statistical power in previous experiments, we preregistered to test transfer only for metacognitive learners.

Participants. We recruited 1100 UK-based participants from Prolific on 19th of July 2024. Participants were pre-screened using the same criteria as in Experiment 1. For the first study of this two-study paradigm, we paid participants £3.78 (US\$4.80) for the 28-minute study (base rate: \$4, bonus payment for passing the attention check: \$0.80). For the second study, which took five minutes, we paid participants £1 (US\$1.27).

A total of 1100 people took part in the first study and 811 fully completed both studies. As preregistered, of those who took part in both studies, we excluded some for failing the attention check ($N = 66$), and those of the remaining participants who indicated that they had participated in a study by any of the experimenters before ($N = 18$). While some participants started the second part directly after the first (23% of participants had a time gap of less than 10 minutes), the majority of our participants had considerable gap between the two studies (62% of participants had a time gap of more than one hour and 45% had a time gap of more than two hours).

Our final sample size was $N = 727$ ($M_{\text{age}} = 39.43$, $SD_{\text{age}} = 13.15$, $N_{\text{female}} = 366$, $N_{\text{male}} = 361$).

Materials and Procedure. The experiment comprised two separate tasks on Prolific, which we will call Experiments 2a and 2b. Experiment 2a comprised the moral learning paradigm from Experiment 1 followed by a few exploratory measures described in the supplementary materials.

Recruitment Procedure. Experiment 2a and Experiment 2b were two separate on-line studies run by different researchers from different institutions. Experiment 2a was posted from the Prolific account of a co-author at UCLA as a task called ‘Moral decision making study’. This task was described as a study run by a researcher at the University of California, Los Angeles; it used the consent form of an IRB protocol issued by UCLA. Experiment 2b was posted with a UCL Prolific account as a task called ‘Donation decisions’. This task was described as a study run by a researcher from University College London and used the consent form from an ethical approval protocol issued by UCL. Therefore, Experiment 2 can be viewed as two independent studies that were later joined together for a cross-study analysis.

The recruitment for Experiment 2b started about half an hour after the start of Experiment 2a and remained open for about 12 hours. Experiment 2b was *only* visible to workers on Prolific who had completed Experiment 2a, but it was impossible for them to know this. Workers who had completed Experiment 2a simply received an email from Prolific inviting them to Experiment 2b (as is standard recruitment procedure) without any further information.

Thus, from the participants’ perspective, Experiment 2a and Experiment 2b were unrelated. We confirmed this assumption by explicitly asking participants at the end of the second study whether they had ever before participated in a study ‘conducted by any of the same experimenters before’;³ only 2.42% said yes, and we excluded these participants from the analysis.

Transfer to New Study. Experiment 2b included the ‘Human Challenge Trials’ and ‘Animal Research’ donation vignettes from Experiment 1. Participants made these allocations knowing that the experimenters would execute the decision of one randomly chosen participant for one randomly chosen pair of charities.⁴

We excluded the ‘Doctor’ vignette from this study because a pilot study con-

³We added the following clarification below the question: ‘Note that this refers to whether you have participated in other studies by the same specific researcher. Do not tick yes if you have only participated in other studies from University College London run by different researchers. This is also not an attention check and your response would not affect your pay, so please answer honestly.’)

⁴We later donated £160 to 1Day Sooner and £40 to Medical Research Foundation on June 28th, 2024.

ducted after Experiment 1 revealed that, contrary to our assumptions, participants neither considered sending doctors to crisis areas to be the option endorsed by CBR, nor did they think that donating to the homeless healthcare charity was the option endorsed by moral rules (see supplementary information).

Data Analysis. We used the same analytical approach as in Experiment 1.

Results

Computational Modeling Results Indicate (Model-Based) Metacognitive Learning. Replicating the previous modelling results, we again found that similar proportions of participants were best explained by each model (see Table 6.1) and each type of learning mechanism (see Table 6.2).

Table 6.1: Cognitive Modeling Results Showing the Proportions of Participants that Relied on Each Type of Learning in Experiments 1-2.

Model			Expected Frequency $\mathbb{E}[f Y]$			
			Expt 1		Expt 2	
			CBR	Rule	CBR	Rule
Model-Based		MB-M	86.97%	31.81%	74.08%	40.24%
Model-Free	Metacognitive	MF-M	3.32%	4.06%	8.26%	1.26%
Constant		C-M	2.53%	3.45%	3.59%	5.55%
Model-Based		MB-B	4.41%	56.17%	9.32%	49.53%
Model-Free	Behavior	MF-B	2.18%	3.82%	2.90%	2.59%
Constant		C-B	0.59%	0.69%	1.85%	0.82%

Note. ‘CBR’ denotes the CBR Success condition, and ‘Rule’ denotes the Rule Success condition. The largest proportions in each condition are highlighted in bold.

Metacognitive Learning Transfers to A Different Experiment. As shown in Figure 6.6, Experiment 2 replicated the transfer effects from Experiments 1. The effect was replicated across experiments and an average delay of about two hours (Mean = 121 minutes, Median = 99 minutes, range = 0.25 to 561 minutes).

We found that evidence for metacognitive learning significantly moderated the effect of the experimental manipulation on all measures of transfer (OUS Sacrificial Harm Subscale, $b = 0.81, t(723) = 7.79, p < .001$; DCS Deontology Subscale, $b = -0.62, t(723) = 4.82, p < .001$; Donations,⁵ $b = 8.69, t(723) = 4.06, p < .001$).

⁵Note that these are the results for the ‘Human Challenge Trials’ and ‘Animal Testing’ vignettes;

Table 6.2: Cognitive Modeling Results Showing the Proportions of Participants that Relied on Each Learning Mechanism in Experiments 1-2.

Learning Mechanisms	Expt 1				Expt 2			
	CBR		Rule		CBR		Rule	
	$E[f Y]$	ϕ	$E[f Y]$	ϕ	$E[f Y]$	ϕ	$E[f Y]$	ϕ
Metacognitive Learning	92.5%	1	35.8%	0.06	85.5%	1	41.9%	0.26
Behavioral Learning	5%	0	60.8%	0.94	11.1%	0	53.5%	0.74
No Learning	2.5%	0	3.4%	0	3.3%	0	4.6%	0

Note. ‘CBR’ denotes the CBR Success condition, and ‘Rule’ denotes the Rule Success condition. The exceedance probability ϕ of a given model family is the probability that the proportion of participants best explained by a model from that family is greater than for any of the alternative model families. The largest proportions in each condition are highlighted in bold.

Further, when including evidence for metacognitive learning as a covariate, we found a significant main effect of condition on all measures of transfer (OUS Sacrificial Harm Subscale, $b = 2.16, t(723) = 8.77, p < .001$; DCS Deontology Subscale, $b = -1.64, t(723) = 5.33, p < .001$; Donations, $b = 20.54, t(723) = 4.04, p < .001$).

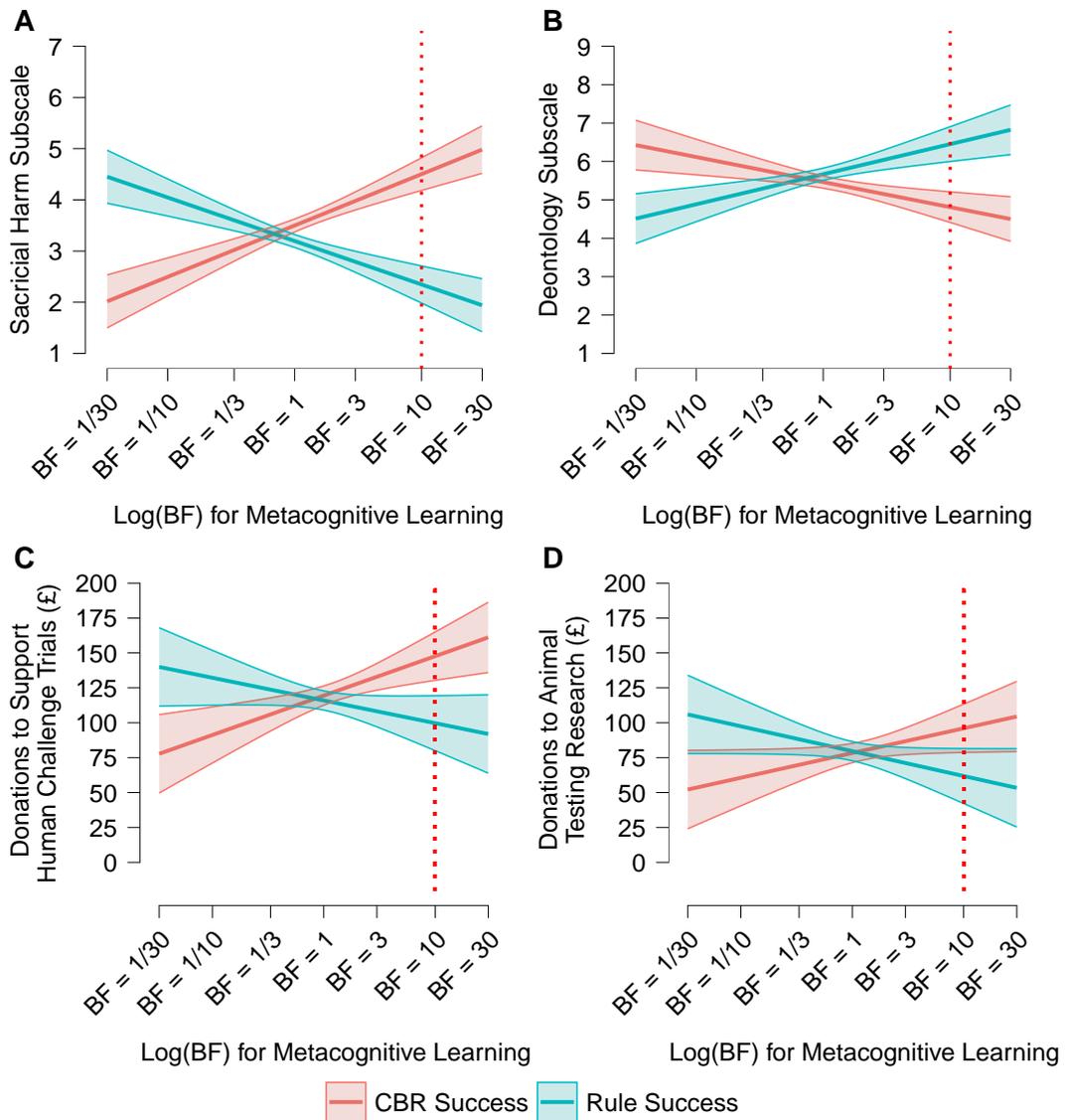
This evidence of transfer to a new experiment, which to participants appeared to be conducted by different researchers, rules out demand characteristics as an alternative explanation of the findings from Experiments 1. Moreover, as detailed in the supplementary information, Experiment 2 also ruled out the alternative explanation that the learning and transfer effects are due to changes in risk aversion.

Individual Differences in Perceived Real-World Relevance and Engagement

Predict Metacognitive Learning (Exploratory). To understand why some participants engaged in metacognitive learning whereas others did not, we measured how they perceived the moral learning paradigm. In brief, we found evidence for metacognitive learning was predicted by participants taking the task seriously, believing that the task allowed them to learn how to make better decisions in the real world, experiencing an emotional response to the outcomes, and perceiving the outcomes as plausible, informative about the real world, and a good reflection of whether they made the right decision (all $p < .02$). Unfortunately, we found no

as preregistered, we removed the donation vignette about sending doctors to crisis zones due to methodological considerations (see Methods Section 6.1.5).

Figure 6.6: Metacognitive Learning Transfers to Measures of Moral Convictions and Donation Decisions in Experiment 2.



Note. Panel A shows responses to the Sacrificial Harm Subscale from the Oxford Utilitarianism Scale. Panel B shows responses to the Deontology Subscale from the Deontological-Consequentialist Scale. Panel C shows the amount of money donated to the charity advocating for human challenge trials. Panel D shows the amount of money donated to a breast cancer research charity that uses animal research. Each plot compares responses between the CBR Success and Rule Success conditions as a function of the amount of evidence the participants' responses in the moral learning paradigm provided for metacognitive learning (BF). BF values > 1 indicate evidence for metacognitive learning. The red dotted line at BF = 10 indicates strong evidence of metacognitive learning; this is where the main effect of the experimental condition was tested. The confidence bands indicate 95% confidence level. See supplementary information for a smoothed conditional means version of the plot.

evidence that any of these factors explained why metacognitive learning was more prevalent in the CBR Success condition than in the Rule Success condition (all $p > .10$).

6.1.6 Discussion

Across two experiments, participants consistently adapted their moral decision strategies based on the consequences of their past choices. When relying on cost-benefit reasoning (CBR) led to better outcomes, participants learned to override strict moral rules in favor of maximizing the greater good. In contrast, when CBR resulted in worse outcomes, participants shifted toward following moral rules instead. This learning not only emerged rapidly but also transferred to a donation decision and scale measures in a different experiment, which participants thought was administered by different experimenters. We only observed this transfer for those participants who showed evidence for metacognitive learning. Together, these findings suggest that meta-control in moral decision making is shaped by fast, adaptive learning from the outcomes of previous decisions.

Although moral learning was driven by consequences, it did not always direct people to rely more on CBR. Instead, when following moral rules had previously produced better outcomes, people learned to rely more on rules. More generally, our findings suggest that learning from consequences aligns decision making with global consequentialism (Ord, 2009), the view that one should adopt whatever decision strategy—whether rule-based, reasoning-based, or virtue-based—produces the best outcomes. Ironically, those who insist on following moral rules ‘regardless of the consequences’ may have reached this position through learning from consequences themselves. In this sense, everyone could be seen as a consequentialist learner, regardless of the moral principles they endorse.

This perspective also sheds light on why people disagree about moral dilemmas. Life experience plays a crucial role: some individuals may have learned that rigidly following rules leads to worse outcomes than flexible, cost-benefit reasoning; others may have found that attempts to outsmart moral rules often backfire. If moral disagreement stems partly from different life experiences, then sharing these

experiences—and learning from each other—may help bridge moral divides.

Although all of our experiments focused on moral decision making, our finding that adaptive metacognitive learning from the consequences of past decisions shapes reliance on different decision strategies might also apply to judgment and decision making more generally. Converging evidence for adaptive metacognitive learning in domains ranging from financial decision making (Erev & Barron, 2005; Lieder & Griffiths, 2017; Rieskamp & Otto, 2006), and cognitive control (Lieder, Shenhav, et al., 2018), and planning (Callaway et al., 2022; He & Lieder, 2023; He et al., 2021; Jain et al., 2019), as well as problem-solving and mental arithmetic (Lieder & Griffiths, 2017) supports this generalisation.

Our results also challenge previous approaches to moral psychology that equated normative theories of morality with decision strategies. As we argued in the introduction of this article, conceptually, deontology and utilitarianism are different from reliance on rules versus CBR, even though they are sometimes equated in the literature. The former are ethical theories that tell us what we should value, whereas the latter are decision strategies that can be used to achieve outcomes that are consistent with those values. In line with research that shows that simpler heuristics can lead to better consequences in certain environments, in the real world, relying on rules may sometimes lead to better consequences than CBR for various reasons (e.g., increased accuracy due to the bias-variance trade-off, Brighton and Gigerenzer, 2012; Gigerenzer and Brighton, 2009; lower cost of computation, Lieder and Griffiths, 2017; and increasing trust, Gigerenzer, 2008). Below, we discuss two findings showing that measures previously considered to measure reliance on different ethical theories may, in fact, measure reliance on different decision strategies.

First, existing self-report measures of deontology versus utilitarianism may actually measure reliance on the specific decision strategies of following rules versus CBR. Our results support this conclusion because (1) learning from consequences can increase people's scores on the Deontology Subscale of the Deontological-Consequentialist Scale (Mata et al., 2022) and decrease their scores on the Oxford Utilitarianism Scale Sacrificial Harm Subscale (Kahane et al., 2018) and (2)

interpersonal differences in these scales were unrelated to evidence of metacognitive moral learning from the consequences of past decisions. If the DCS Deontology Subscale actually measured deontology, we would expect that participants who score higher would be less driven by outcomes and more by the intrinsic rightness of the action and, therefore, show less learning. Instead, we found that participants' scores on this scale were unrelated to how much they engaged in metacognitive moral learning from consequences. Moreover, learning from consequences changed participants' scores on the DCS Deontology Subscale. This suggests that it measures reliance on rules rather than on deontology, and that reliance on rules is more receptive to learning than previously thought, given that these scales are often considered to measure stable traits (Kahane et al., 2018). A similar argument may apply to the OUS Sacrificial Harm Subscale, although the case here is somewhat weaker because this scale claims to only measure one specific component of utilitarian psychology (sacrificial harm).

Second, a participant's 'deontological' or 'utilitarian' choices in moral dilemmas do not necessarily demonstrate that the participant is deontologist or utilitarian. Instead, those choices should be interpreted more cautiously as being consistent with following moral rules or CBR. If we had used participants' decisions in moral dilemmas to measure deontology and utilitarianism, we would have concluded that around 50% of people are deontologist (the proportion that chose the rule option in the first trial), rather than the much lower proportion of people that showed no learning from outcomes (around 5%). This article thereby adds to the existing literature challenging the use of sacrificial dilemmas to measure utilitarian versus deontological decision making (Kahane, 2015; Kahane et al., 2015), and offers an alternative interpretation of these choices in terms of decision strategies.

Our theory also raises the question of how to compare the learning signals from different ethical theories and whether it is possible to have a utility function that is agnostic about which theories people use. Our paradigm is able to capture learning broadly for ethical theories that take consequences into account. This is because we classified the outcomes in our paradigm as good or bad based on partic-

ipants' own evaluations and also use these evaluations for our computational models. This sidesteps the question of how people determine what is a morally good or bad outcome.⁶ As for deontological theories, the intrinsic rightness of certain actions determines their goodness/badness regardless of their outcomes. Therefore, we would not expect that people who strictly follow deontology would learn in our paradigm, as they would not learn from outcomes. In line with this, our theory explicitly acknowledges that moral evaluations also depend on other factors, such as moral intuitions about the chosen action itself (see Figure 6.1, particularly the arrow going directly from decision to moral evaluation). Consistent with these assumptions, we did indeed find that a small proportion of participants did *not* learn from the consequences of their decisions in our task.

Our finding that metacognitive learning was consistently more prevalent in the CBR Success condition than in the Rule Success condition raises the question of which experiences and situational factors trigger versus inhibit metacognitive moral learning. We investigated this question through a series of exploratory analyses reported in the supplementary information. These analyses identified several factors that predict increased versus decreased metacognitive moral learning, including taking the moral dilemmas seriously, the perceived plausibility of the outcomes, the emotional experience of the outcomes, the perceived informativeness of the outcomes and their relevance to the real world, and the perceived utility of learning. However, we also found that none of those factors differed significantly between the CBR Success and Rule Success conditions.

In principle, less learning in the Rule Success condition could have occurred because the majority of our participants already relied on rules in the first dilemma. However, we consistently found that for the first dilemma, around half of the participants chose the rule option, and the other half chose the CBR option in both conditions (Experiment 1: 55%; Experiment 2: 54%). Therefore, it seems unlikely that they had a stronger prior that one option would work better over the other.

⁶The modelling approach we used does not require any integration between the utility functions of different moral theories because we only model intra-individual learning based on a participant's own utility function (rather than trade-offs between the utility functions of different participants).

These results suggest that it is unlikely that the difference between the amount of systematic change between the two conditions results from unintentional differences between conditions in the paradigm. Instead, it might reflect an inherent difference between CBR versus following moral rules: while there is only a single CBR strategy, there are a vast number of different rules one could learn to follow. Our paradigm reflected this reality: the pertinent moral rule(s) differed across the 13 dilemmas. In the CBR Success condition, learning was easier because participants only had to learn about the high effectiveness of CBR, and the decision strategy of CBR is more generally applicable in different contexts than any particular moral rule. By contrast, in the Rule Success condition, metacognitive learning could either guide participants to rely more on rules in general (i.e., relying on rules leads to better outcomes), or to rely more on specific rules (e.g., ‘tell the truth’ and ‘do not kill’). In our experiments, the pertinent rule differed across dilemmas. Therefore, participants who learned about specific rules observed less evidence for the effectiveness of any one rule. Moreover, even when those participants learned to rely more on one of the rules that led to good outcomes, this learning did not necessarily show in subsequent dilemmas where the pertinent moral rules were different. Future research could test this interpretation by conducting experiments in which a salient moral rule is held constant.

Our findings also raise the question of why we found a more consistent effect on participants’ choices (consistent across both experiments) compared to the effect on their moral judgments (for which the effect was only significant in some of the experiments). We conducted a cross-study analysis across all four experiments presented in Maier, Cheung, and Lieder (2024) which showed evidence for an overall effect on judgments and no statistically significant evidence for moderation by Experiment or Condition (Rule Success vs. CBR Success; see supplementary information). However, this still raises the question of why the effect on judgments was weaker than on choices. One possible explanation for the stronger effect of experience on choices is that when giving a judgment of moral rightness, (some) participants might have interpreted the question (‘How morally right is it for you to

[action under consideration]?’) as asking solely about the intrinsic rightness of the action regardless of its consequences in the specific situation. Prior research shows that these types of moral judgments often involve different considerations compared to decisions that would imply reduced learning. For instance, moral judgments tend to be driven more by reputational concerns (Batson & Thompson, 2001). Given that social incentives strongly favor the expression of deontological over utilitarian convictions (Everett et al., 2018), one might expect that people’s ratings are always biased toward deontological principles independent of the anticipated consequences. Future work could explore this by developing a scale that includes different questions, some of which are focused more on the action and others that are focused more on the outcomes, and validating those questions against behavioural measures.

In addition to these future directions, our findings also open up several other avenues for future research. Our new paradigm enables rigorous experiments on moral learning from consequences and lays the groundwork for these follow-up studies. First, our demonstration of metacognitive moral learning raises the question of what the underlying mechanisms are. We have taken a first step towards developing and comparing models of model-free and model-based metacognitive moral learning. Our observation that metacognitive moral learning appears to be more model-based than model-free is consistent with a long series of findings suggesting that model-based learning contributes to many instances of learning that were once assumed to be purely driven by simple model-free RL (e.g., Courville et al., 2006; Daw et al., 2011; Huh et al., 2009; Tolman, 1948). However, our experiments were not optimized for this comparison, and our models also differed along another dimension. That is, the model-free model learns from continuous moral evaluations, whereas the model-based model learns only about the probabilities of binary events (good vs. bad). To address this limitation, the next section of this chapter presents an extension of the two-step task to moral decisions contrasting different decision strategies.

Second, it remains unclear which types of people are more likely to engage in

metacognitive moral learning. Although several aspects of people's perception of our task predicted metacognitive learning, we did not find any relationships with stable individual differences (see supplementary information).⁷

Third, follow-up research could test how stable the moral learning induced by our paradigm is over time. Experiment 2 showed that the effects of learning are not fleeting, as the transfer effects were observed after an average time delay of about two hours. However, considering that the two experiments were still conducted relatively close together in time, it would be necessary to implement these studies with a larger time delay to draw stronger conclusions about how long these effects last.

Fourth, future research could explore the effects of variations in the reinforcement schedule. In the current experiment, we focused on a simple reinforcement schedule, where the rule/CBR options always/never led to success. The reasons for this choice were mostly pragmatic: Our task has relatively few trials compared to other reinforcement learning tasks, and it is difficult to increase the number of trials much more without making the task too long. Therefore, the current schedule achieves the strongest learning signal, given the small number of trials in our experiment. One straightforward modification is to introduce probabilistic rewards (e.g., CBR leads to success 80% rather than 100% of the time). In line with research on intermittent conditioning (Skinner, 1957), probabilistic reinforcement may lead to stronger behaviour maintenance once a given level of behaviour is reached and would, therefore, also be valuable for future work aimed at probing the temporal stability of moral learning.

Fifth, future research should explore different learning signals. Although our experiments focused on learning from consequences, in real-world situations where consequences are unobserved, delayed, or ambiguous, other factors, such as social considerations (Gawronski et al., 2017a, 2023; Kleiman-Weiner et al., 2017; Nichols, 2022), might have a stronger influence on the moral evaluations people learn from. Our paradigm could be adapted to use different kinds of outcomes.

⁷Except for a barely statistically significant association with open-minded thinking about evidence.

Finally, future research should investigate what role metacognitive moral learning plays in moral development and moral learning in the real world. The real-world context that is most similar to our experimental paradigm is learning from stories. The stories we tell our children often teach moral lessons via the fictitious consequences of the protagonists' moral decisions, and so do some of the novels and movies we read and watch. Some teach us that overriding moral rules for anticipated benefits (CBR) leads to good consequences (e.g., *Robin Hood*, *The Imitation Game*). Many others dwell on the tragic consequences of being swayed by the anticipated benefits (CBR) of breaking a moral rule (e.g., *Les Misérables*, *The Mist*, *Minority Report*). The learning we demonstrated in our experiments likely occurs when people encounter such stories. In our experiment, the evidence for metacognitive moral learning was strongest when participants perceived the outcomes to be highly realistic (see supplementary information). This suggests that metacognitive moral learning might be even more powerful for real moral decisions with real consequences.

It has often been argued that human morality is fallible and that people are often swayed by morally irrelevant details (Greene, 2013). While this may be true of people's decisions in traditional philosophy thought experiments, our experiments offer a more optimistic perspective: When people experience the outcomes of their moral decisions, they learn to adopt decision strategies that benefit the greater good. Thus, with sufficient experience, people's morality can become adaptive. From the perspective of ecological rationality (Todd & Gigerenzer, 2012), there is hope that this learning mechanism might tailor human morality to the demands of everyday life (Gigerenzer, 2008). While we do not know whether following rules or CBR would lead to better outcomes in real life (and there is likely no domain-general answer to this question), our research shows that in situations where people receive frequent, prompt, and accurate feedback about the consequences of past decisions, their moral decision making might work much better than people's responses to thought experiments might suggest (cf. Kahneman & Klein, 2009).

The human capacity for moral learning demonstrated by our experiments is

a crucial prerequisite for moral progress (Buchanan & Powell, 2018; Schinkel & de Ruyter, 2017). Unlike social learning, which can propagate bias and prejudice (Schultner et al., 2024), moral learning from the consequences of past decisions can ground people's subjective sense of right and wrong in the objective reality of what alleviates versus causes suffering and what promotes versus reduces well-being (Gibbs, 2019). Some argue that moral progress has been too slow, leaving common morality unprepared for some of the biggest moral problems of the 21st century (Greene, 2002). As an optimistic counterpoint, our findings suggest that when people observe the consequences of their decisions, moral learning can be fast and adaptive.

Reference for Journal Article: Maier, M.*, Cheung, V.*, & Lieder, F. (in press). Learning from Outcomes Shapes Reliance on Moral Rules versus Cost-Benefit Reasoning. *Nature Human Behaviour*. <https://doi.org/10.31234/osf.io/gjf3h>

Data, Code, and Preregistration Availability: All data, code, and preregistrations are available at <https://osf.io/4up5z/>.

Supplemental Information: The supplemental information is available at https://osf.io/preprints/psyarxiv/gjf3h_v2 (as part of the preprint file, starting at page 47).

Author contributions: All authors were involved in the conceptualization and design of the experiments. V.C. and M.M. created the materials with feedback from F.L., programmed the experiments, and collected the data. M.M. preregistered the experiments, conducted data analysis and visualization, and V.C. checked the analysis code. All authors contributed to writing, reviewing, and editing the manuscript. F.L. supervised the project.

6.2 Disentangling Model-Based and Model-Free Moral Learning

While the previous sections shows that participants engaged in metacognitive moral learning, more evidence is needed to ascertain whether they had done this by constructing a model linking strategies to outcomes (i.e., they engaged in *model-based*

learning), or whether good/bad outcomes reinforced the strategies directly (i.e., they engaged in *model-free learning*). As outlined in the discussion section of the previous experiment, the paradigm used in the first section of this chapter was not optimized for disentangling the two types of learning.

To address this gap in the current understanding of moral learning, we sought to disentangle model-free from model-based moral learning from the consequences of past decisions by developing a new experimental paradigm that is optimized for answering that very question: the moral two-step task. In this section, we present this paradigm, leverage it to conduct an experiment, and analyse the data using computational models of model-free and model-based moral learning.

6.2.1 The Moral Dilemma Two-Step Task

Within the RL framework, the two-step task is the dominant method for dissociating model-based and model-free learning (the task presented in this section combines features from versions proposed by Daw et al., 2011 and Kool et al., 2016). Each trial consists of a sequence of two decision stages. In the first-stage state, participants choose from two options. Each choice leads to one of two second-stage states. Participants then press a button to reveal the second-stage state, which is determined by a probabilistic reward function that changes during the task. The transition from the first-stage choice to the second-stage state is also probabilistic, but the probabilities are fixed during the experiment. Each first-stage choice has a high probability (e.g., 70%) of transitioning to one of the two sets of options (common transition) and a low probability (e.g., 30%) of transitioning to the other set of options (rare transition).

The design of this task ensures that model-free and model-based learning have opposite effects on the probability that the first decision will be repeated after a rare transition led to a reward. Model-free learning increases the probability that the first decision will be repeated, but model-based learning reduces the probability that it will be repeated. This is because model-free learning, which relies on an action's average consequences in the past, would disregard the transition type and simply repeat the successful action in the first-stage state which ultimately led to

a good outcome. In contrast, model-based learning, which uses a model of the environment, recognizes the second-stage state that led to a good outcome came as a result of a rare transition of the chosen first-stage option. In other words, it recognizes that the good outcome is the result of a common transition of the *unchosen* option, and increases the probability of choosing this option in the future.

Some prior studies have applied the two-step task to tasks involving moral judgment or behaviour. For instance, Patil et al. (2020) used this paradigm to investigate whether model-based learning is associated with utilitarian reasoning, and P. L. Lockwood et al. (2020) used it to investigate whether learning of harm aversion is model-free. Both studies adapted the paradigm to the context of moral decision making by changing the reward state to inducing or avoiding physical harm (loud noise or electric shock) to another person. Apart from this, the decision states were very similar to the non-moral versions of the task: participants' decisions in the first-stage state (e.g., which of symbols to select) had no resemblance to moral decisions.

These types of first-stage decisions do not capture relevant features of making a moral decision. Moral decision making usually involves choosing between two contrasting options in a dilemma (e.g., tell the truth vs. lie but help more people). Therefore, we developed a novel version of the two-step task in the context of moral decision making. In this task, the first stage confronts people with a moral dilemma between following moral rules versus CBR.

In the moral dilemma two-step task, participants are faced with a moral dilemma in the first-stage state with two conflicting choices: one endorsed by following moral rules and one endorsed by CBR. Each choice leads to a common (70%) versus rare (30%) transition into a second-stage state. The two possible second-stage states leads to good versus bad outcomes probabilistically. We measured participants' first-stage choices, namely the probability of choosing the same option as they did in the previous trial (i.e., 'stay probability').

Following the logic of Daw et al. (2011), in our task, if moral learning were model-free, participants would increase their reliance on CBR/rules whenever it led

to a good outcome, regardless of the transition type. If moral learning is model-based, then participants would increase their reliance on CBR/rules only if the good outcome resulted from a common transition but decrease it if the good outcome resulted from a rare transition.

6.2.2 Experiment

Method

We preregistered our experiment and data analysis on AsPredicted (#185758).

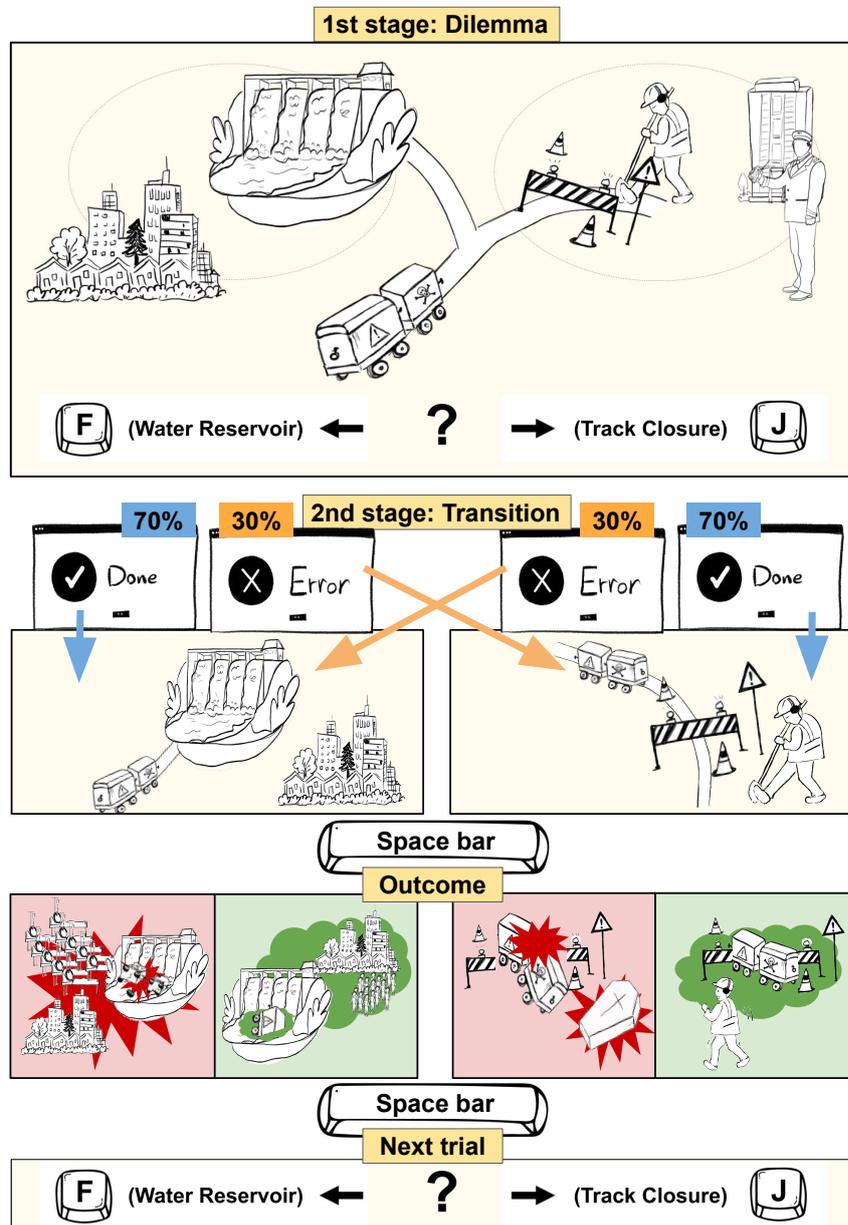
Participants. This experiment received ethical approval from the Office of the Human Research Protection Program at University of California, Los Angeles, under protocol number IRB#23-001436.

We recruited 150 participants from a U.K. and U.S. sample on Prolific on August 8, 2024, based on an a priori power analysis which indicated that 100 participants are required. The median duration for the study was 44 minutes. We excluded participants whose responses were not fully recorded due to a technical issue ($N = 8$). As preregistered, we also excluded participants who did not pass the second attempt at a comprehension check after reading the task instructions ($N = 8$). Of those who passed the comprehension check, we excluded those who failed two or more attention checks during the main task ($N = 4$). Our final sample size was $N = 130$ ($M_{\text{age}} = 37.23$, $SD_{\text{age}} = 11.96$, $N_{\text{female}} = 63$, $N_{\text{male}} = 67$).

Design and Materials. Participants made 125 repeated decisions. As shown in Figure 6.7, in the first-stage decision state, we presented a trolley-type moral dilemma. Participants imagined themselves as an employee of a railroad company whose job is to oversee railway junctions where two tracks diverge. A runaway wagon containing dangerous pathogens and explosive materials is quickly approaching the junction. The participant cannot stop the wagon, but must decide whether to direct it to the left or right track. Participants do this by pressing F (for the left track) or J (for the right track) on their keyboard. Unlike classic trolley-type dilemmas, where one must decide between acting to divert the wagon to an alternative track or doing nothing and letting it stay on the default track, we designed the decision to

be between switching to a left or a right track to avoid confounding one choice with action and the other with inaction (see Crone and Laham, 2017).

Figure 6.7: The Moral Dilemma Two-Step Task



Note. In the moral dilemma two-step task, the first-stage decision state had two conflicting choices: one endorsed by following moral rules (saving your colleague by diverting the wagon to the left track, risking the contamination of a water reservoir causing a disease with 2% mortality rate), and one endorsed by CBR (saving more lives by diverting the wagon to the right track, which is prohibited by authorities and risks killing your colleague). Each of these choices led to a common (70%; wagon moving as directed) versus rare (30%; wagon turning in the opposite direction due to a system error) transition into a second-stage state. The second-stage state then led to good versus bad outcomes probabilistically.

In all trials, the left track leads to a water reservoir that supplies water to a city. If the wagon lands in the reservoir, it would likely explode and leak pathogens into the water, causing a serious disease in about 100 people in the city with a mortality rate of $\sim 2\%$. However, there is a chance that the wagon would remain intact, in which case the water would stay clean. The right track has been closed by the railroad authorities, as indicated by a track closure sign. A worker, who is a colleague of the decision-maker, is conducting maintenance work on the track. If the wagon heads towards him, he would likely be killed. However, there is a chance that he would see the wagon and escape. Participants were also informed that their boss had instructed that railroad company employees have a duty to always obey the track closure signs.

Diverting the wagon to the water reservoir is the option endorsed by moral rules (i.e., the *rule option*), because it follows the moral principle ‘do not kill’, as well as the moral obligation to protect one’s colleagues and obey the rules set by an authority figure. Diverting the wagon to the track closure is the option endorsed by CBR (i.e., the *CBR option*) because fewer people would be sacrificed. We had previously piloted these choices to ensure that they are well-balanced, meaning that around 50% of people would prefer each option. The pilot study ($N = 60$) revealed that $M = 45\%$, 95% CI [32%-58%] preferred the rule option, and $M = 55\%$, 95% CI [42%-68%] preferred the CBR option. Further, to check whether participants indeed perceived diverting the wagon towards the track closure to be the option endorsed by CBR, and diverting the wagon towards the water reservoir as the option endorsed by following moral rules, we asked four questions where participants indicated which track would someone who always (1) followed rules, (2) followed their gut feelings, (3) considered the best consequences, and (4) followed reasoning, would choose. As expected, the majority of participants selected the rule option for questions (1) and (2) (70% & 67%), and the CBR option for questions (3) and (4) (74% & 71%).

Each choice in the first-stage moral dilemma led to a common (70%) versus rare (30%) transition into a second-stage state. As part of the cover story, participants were told that every time they choose which track to divert the wagon to, their

choice is supplied to a computer system that executes their decision; however, this system is unreliable and sometimes mistakenly diverts the wagon to the opposite (unchosen) track. In the task, there was a 30% chance that when participants chose to divert the wagon to one track, it would actually be diverted to the other (i.e., rare transition to the second-stage state). When this happened, they saw a symbol on the screen indicating a system error and that the wagon went to the opposite track. However, most of the time, the computer worked properly (i.e., common transition), and the participant saw a symbol indicating that their order went through successfully and that the wagon went to the specified track (see Figure 6.7).

The second-stage state then led to good versus bad outcomes probabilistically. Either a good or bad outcome could occur regardless of which track the wagon was diverted to. If the wagon goes to the water reservoir, either it remains intact, and nobody is harmed (good outcome), or it explodes, and 100 residents get a disease with a 2% mortality rate (bad outcome). If the wagon goes to the closed track, either the colleague escapes the track, and nobody is harmed (good outcome), or the wagon runs over the colleague and kills him (bad outcome). The reward probabilities of the second-stage choices (i.e., whether the outcome is good or bad) changed according to a Gaussian random walk ($\mu = 0, sd = 0.025$), initialized at .5 in the first trial and with limits imposed on its range (between 0.25 and 0.75).

Procedure. The experiment was programmed on the online platform lab.js. After giving informed consent, participants first read instructions describing the task, including the two choice options, the rare and common transitions, and the positive and negative outcomes that could occur as a result. As part of the instructions, participants completed three rounds of practice trials, with ten trials each. We included three comprehension check questions after the instructions and before the main task. Participants who responded incorrectly to at least one question were required to review the instructions and make a second attempt. Participants who failed any of the comprehension check questions on their second attempt were excluded from analysis.⁸

⁸In the experiment, of those participants who passed the comprehension check questions, 83% passed on their first attempt, and 17% passed on their second attempt.

Participants then took part in the main task, which included 125 trials. We included eight attention checks throughout the task, which asked participants to indicate the outcome of their previous decision.

At the end of the task, as exploratory measures, we asked participants a series of questions about how they understood and interpreted the task. We also asked participants if they thought the changes in the study were random or systematic changes, whether the outcomes of previous trials influenced their decision in the next trial, and how frequently they thought the system error (i.e., rare transitions) occurred.

Lastly, participants answered several questionnaires to investigate which factors explain reliance on model-based versus model-free metacognitive moral learning. We asked participants how morally right it was to direct the trolley toward either the (1) colleague or (2) the reservoir. We also asked them to complete various scales relevant to moral decision making, including the Empathic Concern (EC) and Emotional Empathy (EE) Scale (Jordan et al., 2016), The Deontology-Consequentialist Deontology Subscale (Mata et al., 2022), the Sacrificial Harm Subscale of the Oxford Utilitarianism Scale (OUS) (Kahane et al., 2018), and the Cognitive Reflection Test (CRT; Frederick, 2005).

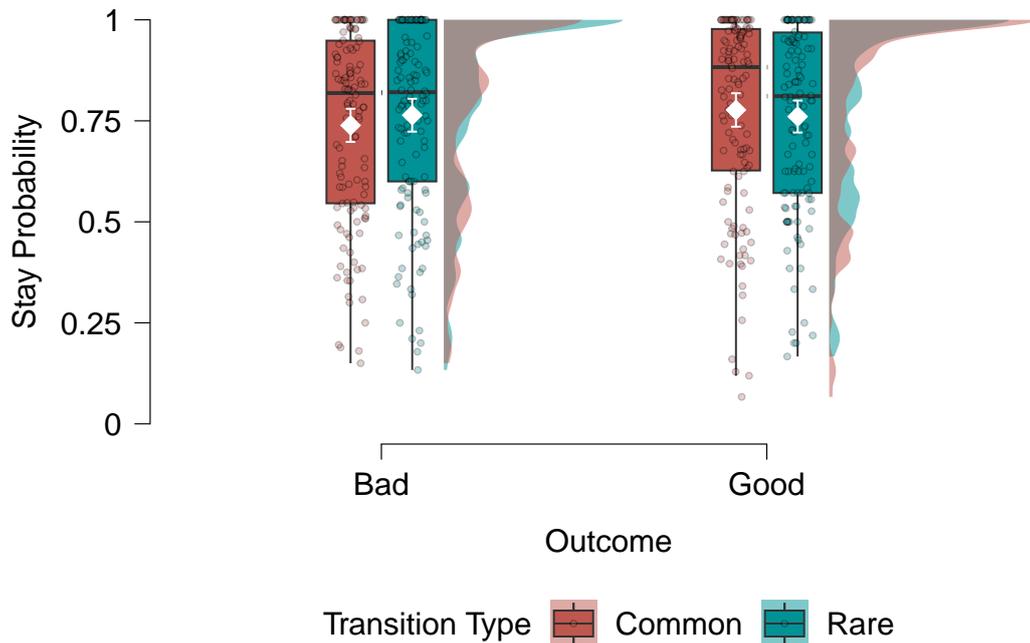
All materials are available on the online repository.

Results

Choice Behavior. On average, participants more often redirected the wagon to the track closure with the colleague (56%, 95% CI [55%, 57%]) than to the water reservoir (44%, 95% CI [43%, 45%]). These proportions were significantly different from 50%, $\chi^2(1) = 226.62$, $p < .001$.

We conducted a logistic mixed effects model using the `afex` package (Singmann et al., 2022) in R to analyse the probability that participants chose the same first-stage choice action as they had chosen in the previous trial (*stay probability*) as a function of outcome (good vs. bad), transition type (rare vs. common), and their interaction. For each participant, the model included random slopes

Figure 6.8: Participants' Choices Reflect Both Model-Based and Model-Free Learning



Note. Stay probability was higher for common transitions compared to rare transitions for good outcomes, and lower for common transitions compared to rare transitions for bad outcomes. Mean and 95% CIs are indicated in white. Boxplots indicate median with IQR, and whiskers are constructed using 1.5 times the IQR.

for transition type, outcome, and their interaction.

Model-based learning from a good outcome increases the stay probability after a common transition but reduces it after a rare transition (and vice versa after a bad outcome). Therefore, an interaction of outcome and transition type would be evidence for model-based learning. Model-free learning would imply increasing the probability of actions that lead to a good outcome and reducing the probability of actions that lead to a bad outcome, independent of whether they followed a common or rare transition. Therefore, a main effect of outcome on the probability of repeating the same action would be evidence for model-free learning.

As seen in Figure 6.8, we found a significant main effect of outcome on stay probability: participants repeated their choice more often after a good outcome than after a bad outcome, $\chi^2(1) = 5.19$, $p = .023$, $OR : 1.08$, indicating model-free learning based on our preregistered predictions.

Moreover, we found a significant interaction between transition type and outcome, $\chi^2(1) = 7.87$, $p = .005$, $OR : 1.09$, indicating model-based learning.

Specifically, when the outcome resulted from a rare transition, we found no evidence that the stay probability was affected by whether the outcome was good ($M = 0.886$, 95% CI [0.837,0.921]) or bad ($M = 0.887$, 95% CI [0.839,0.922]), $OR = 0.99$, $z = -0.13$, $SE = 0.10$, $p = 0.894$. However, when the outcome resulted from a common transition, the stay probability was higher after good outcomes ($M = 0.902$, 95% CI [0.861,0.932]) than after bad outcomes ($M = 0.870$, 95% CI [0.818,0.908]), $OR = 1.37$, $z = 4.17$, $SE = 0.10$, $p < .001$.

All results, including exploratory analysis and manipulation check responses, are available on the online repository.

6.2.3 Computational Modeling

To assess model-free and model-based learning on the level of individual participants, we implemented RL models of model-free and model-based moral learning.

Model Specification and Implementation

Model-Free Learning. Model-free learning is represented by a simple Q -learning model where people directly learn about the associations between the first-stage choice (direct the wagon towards the reservoir vs. the colleague) and the corresponding second-stage outcome. Specifically, this model learns the anticipated moral values of relying on CBR $Q^{\text{MF}}(\text{CBR})$ versus rules $Q^{\text{MF}}(\text{rules})$. These Q -values are updated based on the prediction error between the anticipated consequences of relying on CBR/rules and the observed outcome. If the choice in trial t was consistent with CBR (i.e., the trolley was directed toward the colleague), the prediction error would be calculated as:

$$\text{MPE}_t = Q_t^{\text{MF}}(\text{CBR}) - R_t, \quad (6.11)$$

where MPE denotes the moral prediction error, and R denotes the value of the observed consequences. R was defined as 1 when the decision led to good consequences and -1 when it led to bad consequences. The moral prediction error then guides how much the Q -value for reliance on CBR is updated. The size of the

update is controlled by the learning rate α :

$$Q_{t+1}^{\text{MF}}(\text{CBR}) = Q_t^{\text{MF}}(\text{CBR}) - \alpha \times \text{MPE}_t, \quad (6.12)$$

If the participant decided to follow the rule (i.e., the trolley was directed towards the reservoir), the same updating rules are applied to $Q^{\text{MF}}(\text{rules})$.

Model-Based Learning. Either action can lead to one of two states: either the trolley will head toward the reservoir ($s_{\text{reservoir}}$) or the trolley will head toward the colleague ($s_{\text{colleague}}$). We modeled model-based learning as people using a model of how likely either decision strategy is to lead to either state, people learning the expected outcomes of both states ($V_t(s_{\text{reservoir}})$ and $V_t(s_{\text{colleague}})$), and using their mental model and the learned values to reason about whether it is better to rely on CBR versus rules. The values of the two states are updated in the same way as the model-free Q -values using Equations 6.11 and 6.12, except that they assign value to the states rather than decision strategies. Assuming that people's mental model captures the stated transition probabilities (as in Daw et al. (2011) and a reasonable assumption given the practice trials), we can model the estimates people reach by reasoning about the consequences of relying on CBR versus rules as:

$$\begin{aligned} Q_t^{\text{MB}}(\text{CBR}) &= 0.7 \times V_t(s_{\text{colleague}}) + 0.3 \times V_t(s_{\text{reservoir}}), \\ Q_t^{\text{MB}}(\text{rules}) &= 0.3 \times V_t(s_{\text{colleague}}) + 0.7 \times V_t(s_{\text{reservoir}}). \end{aligned} \quad (6.13)$$

Decision Execution. To implement the assumption that the participants may rely on both model-free and model-based learning and that the weight given to each of the learning mechanisms varies between participants, our model's decisions follow a weighted average of the model-based and model-free Q -values:

$$\begin{aligned} Q_t^{\text{total}}(\text{CBR}) &= w \times Q_t^{\text{MB}}(\text{CBR}) + (1 - w) \times Q_t^{\text{MF}}(\text{CBR}), \\ Q_t^{\text{total}}(\text{rules}) &= w \times Q_t^{\text{MB}}(\text{rules}) + (1 - w) \times Q_t^{\text{MF}}(\text{rules}), \end{aligned} \quad (6.14)$$

where w denotes the decision weight of the model-based system, which is estimated for each participant and allows us to quantify how much participants rely on model-

free versus model-based learning.

Finally, the decision mechanism is selected probabilistically according to the softmax function:

$$p_t(\text{CBR}) = \frac{e^{\tau \times Q_t^{\text{total}}(\text{CBR})}}{e^{\tau \times Q_t^{\text{total}}(\text{CBR})} + e^{\tau \times Q_t^{\text{total}}(\text{rules})}} \quad (6.15)$$

Prior Distributions. As a prior on the learning rate α , we use a uniform distribution on the interval $[0, 1]$. This prior reflects the belief that learning rates larger than 1 (which would imply changing one’s belief by more than the prediction error) and learning rates lower than 0 (which would imply learning in the opposite direction of the prediction error) are impossible. As prior on the mixing weight w was also a uniform distribution on the interval $[0, 1]$, which indicates that all weights given to model-free versus model-based learning are considered equally likely a priori. As the prior distribution on the temperature parameter τ , we use the lognormal distribution $\text{lognormal}(0, 1.4)$, which assigns 90% of the prior probability mass to values of τ between $\frac{1}{10}$ and 10. Finally, we use a prior of $\text{Normal}(0, 1)$ for the initial values of V_t and Q^{MF} .

Model Implementation. We implemented the models in `stan` version 2.35.00 and fitted them using `cmdstanr` version 0.8.0.9000. The model was fitted individually to each participant using four MCMC chains and 20000 iterations per chain, 10000 of which were warm-up iterations.

Prior Predictives & Recovery Simulations. Prior predictive checks confirmed that simulating from the model with different mixture weights w given to model-based versus model-free learning recreates the main patterns in the data associated with these learning types (i.e., the main effect of reward, the interaction of transition type and reward, or both). Further, we conducted recovery simulations where we simulated from the model using different mixture weights w and then fitted the model to the simulated data. This model recovered the weights well, indicating that the mixture parameter can, in principle, be estimated from the data.

Results

Descriptive Results. We removed two participants with $\hat{R} > 1.01$ (Vehtari et al., 2021). For all other participants, the model showed good convergence. Overall, the computational modelling results indicate a weak preference for model-free over model-based learning: The average weight of model-based learning was 45%. 57% of participants have a mixture weight w of less than .5 and 26% less than .25 (w ranges from 0 to 1, whereby a value of 0 would indicate only model-free learning took place, and a value of 1 only model-based learning). In contrast, 43% of participants had a w larger than .5 and only 10% larger than .75.

Relationship Between Evidence for Model-Based Learning and Exploratory Measures. Because there was low colinearity between the various self-report measures (all VIF < 1.5), we tested the relationship between these variables and evidence for model-based learning (i.e., the posterior median estimate of w) within a single regression model. This indicated no evidence that any of those scales were associated with evidence for model-based learning (all $p > .05$; see online repository for a complete report of all results).

6.2.4 Discussion

In this section, we developed a novel paradigm – the moral dilemma two-step task – that can be used to dissociate model-based from model-free moral reinforcement learning. This paradigm was a version of Daw et al. (2011)’s two-step task, adapted to the context of a moral dilemma between following moral rules versus CBR (see Figure 6.7). Our results from behavioural data and computational modelling provided converging evidence that the moral dilemma two-step task successfully dissociates between model-free versus model-based moral learning. We found that moral learning from the consequences of previous decisions combines model-free and model-based learning. Our modelling results suggest that, while the weights given to each mechanism differ considerably between participants, most, if not all, individual participants relied on both mechanisms to some extent.

Previous work on models of moral decision making (Cushman, 2013; Crockett,

2013) investigated moral decision making at the behavioural level (i.e., choosing between actions). In these theories, utilitarian morality (CBR) is associated with the model-based decision system (e.g., saving more lives; Patil et al., 2020) and deontological morality with the model-free decision system (e.g., harm aversion; P. L. Lockwood et al., 2020). The co-existence of these two systems necessitates a mechanism that arbitrates between them when they imply conflicting choices. In the previous section, we showed that this arbitration is at least partly accomplished by learning to rely on the system that produced better outcomes in similar previous decisions (metacognitive moral learning). Our findings are consistent with the idea that metacognitive moral learning occurs through both model-free and model-based reinforcement learning.

One alternative explanation for our findings could be that rather than learning about decision mechanisms, people learned about specific behaviours (e.g., whether to press F or J; whether to direct a wagon toward a colleague or a water reservoir). Based on the evidence for metacognitive moral learning in the previous section of this Chapter, which generalised across different dilemmas and moral decision making tasks, it is unlikely that this is the only type of learning that occurred in our experiment. However, this interpretation cannot be entirely ruled out with the data collected in the current version of the moral dilemma two-step task. We are currently working on follow-up experiments that directly delineate metacognitive and behavioural learning within the two-step task by using different decision scenarios in different trials (and testing the transfer between them) and adding transfer measures after the experiment.

Another interesting avenue for future work is to investigate which traits explain how much people rely on model-based versus model-free learning. It is often difficult to relate questionnaire measures to responses in behavioural tasks, likely because the reliability of tasks can be low when they are used to measure individual differences (Pedroni et al., 2017). In line with this, we did not find a relationship between different exploratory scales and individual level evidence for model-based versus model-free learning. In the current study, we calibrated the statistical power

to be able to detect model-based and model-free learning in the choice data rather than for relating evidence for either learning type to questionnaire measures. Given the low reliability of tasks as individual difference measures, the study likely had much lower power for the latter purpose. We are planning follow-up studies to test these relationships with increased sample size and number of trials to allow for a more powerful test of determinants of model-based versus model-free moral learning.

The moral dilemma two-step task allows researchers to study and dissociate model-free and model-based moral learning from consequences, making it a valuable tool for research on moral learning from outcomes. Further, by modifying the learning signal our paradigm could straightforwardly be adapted to investigate other types of moral learning, such as learning from social feedback, creating applicability in a variety of research areas in moral psychology and beyond.

At a broader level, understanding the mechanisms underlying metacognitive moral learning informs our understanding of moral development and moral progress. This improved understanding may, in turn, strengthen the scientific foundations for fostering moral development, and help inform researchers and educators in developing interventions that empower individuals to translate their ethical commitments into action.

Reference for Journal Article: Tahmasebi, Z.*, Maier, M.*, Cheung, V.*, Cushman, F., & Lieder, F. (in press at *Proceedings of the Annual Meeting of the Cognitive Science Society*). Disentangling Model-Based and Model-Free Moral Learning. https://doi.org/10.31234/osf.io/azhnq_v1

Data, Code, and Preregistration Availability: All data and code are available at <https://osf.io/g6smf/>. The preregistration is available at <https://aspredicted.org/6t37-z5d3.pdf>

Author contributions: All authors were involved in the conceptualization and design of the experiments. Z.T. created the materials and programmed the experiments with feedback from the other authors. M.M. and Z.T. conducted data analysis. M.M.

created visualizations and conducted the computational modelling analysis. Z.T., V.C., & M.M wrote the first draft of the manuscript. All authors contributed to writing, reviewing, and editing the manuscript. F.L. supervised the project.

Chapter 7

General Discussion

In this thesis, I have addressed several crucial topics related to moral and prosocial decision making, ranging from nudging to extinction risk, employing various methods, including publication bias-adjusted meta-analysis, experimental, and computational modelling work. I started this thesis with two examples that show limitations in human moral and prosocial decision making: PlayPump, where altruistically motivated people misallocated resources to a project that had intuitive appeal but was ultimately not cost-effective, and the abolition of slavery, where people's values changed to realise ethical problems with a practice that had been widely accepted at the time. In the next section ('Domain Insights About Moral and Prosocial Decision Making'), I will review what we have learned and how it relates back to the issues of (1) altruistically motivated people misallocating resources and (2) how people learn and change which ethical principles they rely on.

Further, next to the domain insights gained, this thesis also has various methodological implications. First, I highlight the pervasive issue of publication bias in research on moral and prosocial decision making, which leads to substantial overestimation of evidence on various topics, such as Construal Level Theory, the Identifiable Victim Effect, and nudging. This raises the question to what extent we can trust findings in this literature and what practices can be adopted to reduce publication bias. Second, I show how computational modelling can be used to better operationalise theories and to understand behaviour at an individual level. This can be one remedy to mitigate publication bias, as more precise predictions and

theory construction make null findings more informative. After highlighting the substantive and methodological implications of the work, I outline a methodological pipeline that can integrate computational modelling and experimentation with publication bias adjustment.

7.1 Domain Insights About Moral and Prosocial Decision Making

7.1.1 Insights from Publication Bias Adjusted Meta-Analysis

In the first part of the thesis, I show strong publication bias in research on different interventions and theories related to moral and prosocial decision making, in particular, Construal Level Theory (Trope & Liberman, 2010), the Identifiable Victim Effect (Small et al., 2007), and nudging (Thaler & Sunstein, 2009). All of these re-analyses show strong evidence for publication bias: after adjusting for publication bias, the mean effect for all of these phenomena is not robustly distinguishable from zero. Does this, therefore, imply that previous work finding evidence for these phenomena was spurious and they should be removed from the corpus of psychological knowledge? I would argue that the inferences that should be made from these null findings after publication bias adjustment are more complicated, and alternative interpretations are possible.

First, the meta-analyses of nudging and Construal Level Theory show substantial evidence for heterogeneity, even after analysing data within appropriate subgroups. This heterogeneity implies that even when there is no evidence for the mean effect, certain paradigms and interventions within the topic studied by the meta-analysis are likely still effective. This might further be amplified by researchers' preference to test more surprising nudges rather than more obvious nudges with large effects, as the publication system rewards surprising and novel findings. Therefore, especially in the domain of nudging, the academic literature could contain a selection of studies that are less likely to replicate as compared to interventions that would actually be rolled out in a real-world setting.

Second, even to the extent that the meta-analysis does not show evidence for an effect and no evidence for heterogeneity (as was the case for the meta-analysis on the Identifiable Victim Effect), this might merely imply that with the typical measures and manipulations employed the effect cannot be detected rather than that it is not a genuine psychological phenomenon. For instance, in the case of the Identifiable Victim Effect, participants usually see a photo of the identified victim versus statistical information about a group of unidentified victims. Merely seeing a photo of the victim is a relatively weak intervention relative to the types of mechanisms by which a victim would often be identified in real-world decisions (e.g., physically seeing the person). The effect may therefore still apply to certain real-world decisions even if not observed in the lab. However, to the extent that this is the case, the added value of the laboratory research conducted on this topic is questionable.

Overall, these reanalyses, together with other analyses showing strong publication bias in the psychological literature that my collaborators and I conducted, which were beyond the scope of this thesis (Bartoš, Maier, Stanley, & Wagenmakers, 2022; Bartoš et al., 2023, 2024; Maier, Bartoš, & Wagenmakers, 2023), cast scepticism on many seemingly well-established phenomena. While I would stop short of arguing that these phenomena should be discarded for the reasons outlined in the previous two paragraphs, the reanalyses highlight the need for (1) robust methodological practices such as registered reports that avoid publication bias, and (2) better methodologies and theories to understand boundary conditions and heterogeneity (see also the section ‘Where Next? Bridging Publication Bias Adjusted Meta-Analysis and Computational Modelling’).

7.1.2 Insights from Experimental Work

In the second part of my thesis, I turned to empirical work on neglected issues in research on moral and prosocial decision making. In these areas, there was insufficient prior research to make them amenable to the application of bias correction techniques. First, I investigate the method of unit asking and test to what extent it can increase people’s scope sensitivity. Unit Asking is a method whereby participants are first asked how much they are willing to donate to help a single affected

individual, and in a next step, asked how much they are willing to contribute to help a larger group of individuals. Contrary to claims by Hsee et al. (2013), we found that unit asking does not increase sensitivity to scope in a way that people scale their concern with the number of entities affected. Instead, by developing an extension that we term *Sequential Unit Asking* we find that participants increase their willingness to donate proportional to the number of questions asked rather than the number of people.

After the chapter on Unit Asking, I turn to another area where scope neglect or neglect of stakes plays a crucial role: decision making under extinction risk. To do so, I developed a new experimental paradigm that can be used to study these types of risks. Across three experiments studying individual decision making under extinction risk, we find that people are overall closer to the optimal strategies than we had expected; however, they nevertheless show evidence for certain behavioural biases, including insensitivity to the maximum trial number (scope insensitivity) and loss chasing. Finally, we generalise this task to collective decisions and show that people are considerably more risk-taking when playing collectively than when playing individually, though our pilot findings suggest that providing information about the collectively optimal strategy can reduce this risk-taking.

In the last chapter, I turn to the question of moral change and how people learn to decide which moral decision mechanism to rely on. To do so, we first delineate the computational level problem of moral decision making (e.g., which is the best moral theory to rely on) from the algorithmic problem (e.g., which decision mechanisms can best approximate a selected moral theory). We leverage this distinction to design a learning task, where we manipulate the outcomes of people's decisions to either direct them toward cost-benefit reasoning or toward reliance on moral rules. We find that many people adaptively learn to rely on whichever decision mechanism brings about the best consequences, and that this learning generalises to a new experiment by different experimenters (thus ruling out demand effects). These results demonstrate that in the right learning environment, human moral decision making is surprisingly adaptive and can direct people toward decision mechanisms

that usually lead to better consequences. This finding has important implications for the possibility of moral change: If people can observe the consequences of their actions and are free to act as they wish, we would expect that societies can relatively quickly change toward more beneficial moral norms.

Overall, the second part of the thesis offers a somewhat more optimistic perspective on moral and prosocial decision making and ways to improve it. Sequential unit asking effectively increases people's willingness to donate in order to help large groups of others. In the collective extinction gambling task, providing information about the optimal strategy can improve group decision making. Finally, a relatively simple learning intervention can direct people toward moral decision mechanisms that benefit the greater good. This raises the question of what makes these paradigms effective, unlike many of the paradigms for which we found weak evidence at best after adjusting for publication bias in part 1 of the thesis.

One reason is likely to do with the fact that all of the interventions are relatively strong and target the behaviour in question relatively directly. For instance, the moral learning paradigm involves a reinforcement task of around 20 minutes directly targeting moral decision making. In contrast, many interventions in the extant literature may be more indirect by trying to address another psychological capacity and observing downstream effects on the dependent variable (e.g., changing people's climate change perceptions through construal level interventions, Brügger et al., 2016) and relatively small in terms of the amount of time and effort required on the side of participants (e.g., nudges that ask people to sign at the top rather than the bottom of a document).

A second reason is that modelling individual differences allows us to estimate the effect of the intervention only for those participants for whom the manipulation worked as intended. For instance, in the moral learning paradigm, the transfer in the aggregate data is relatively weak; however, for those participants who engaged in meta-cognitive learning, the transfer effects are much stronger. Overall, these findings show that while it is possible to enhance prosocial decision making, there are usually no easy fixes or solutions. Changing moral and prosocial behaviour requires

strong interventions in the same domain as the targeted behaviour, combined with an understanding of the mechanisms behind the interventions and how they play out for different individuals.

7.2 Implications for Methodology and Research Practices

7.2.1 Mitigating Publication Bias

This thesis shows that publication bias is a pervasive issue in research on moral and prosocial decision making. Studies that show evidence for an effect are more likely to be published than studies that do not, which leads to substantial overestimation of effect sizes in the published literature. While this in itself should be no surprise to anyone who has conducted psychological research before, the extent to which publication bias distorts the results is larger than I had expected: there is no evidence for (the mean effect of) some of the most basic theories and phenomena in the field after taking publication bias into account.

This state of affairs highlights the pressing need to adopt publication practices that mitigate bias and researchers' degrees of freedom. One such approach is pre-registration, whereby researchers specify their hypotheses and data analysis plans in advance (e.g., Nosek et al., 2018). However, preregistration is no panacea as reviewers often do not check the preregistration and authors deviate without disclosing it (van den Akker, Bakker, et al., 2024; van den Akker, van Assen, et al., 2024). Further, researchers may still choose not to publish their preregistered study if the results are not to their liking.

Therefore, I believe that where the experimental design is suitable for doing so, the field should ideally go one step beyond preregistration and adopt registered reports (Chambers, 2013; Chambers et al., 2015). Registered reports receive peer review based on introduction and methods before the results are known. If the introduction and methods are sound, the paper receives in principle acceptance, meaning that it is going to be published, independent of the results. Unlike preregistration,

registered reports remove incentive for non-publication of null findings and questionable research practices, as researchers know that their paper will be published as long as they follow the methods that they outlined in the stage 1 submission. In light of these benefits, Chapter 6 (Section ‘Collective Decision Making Under Extinction Risk’) included a pilot study for a paper that I submitted as a registered report.

In addition to the adoption of registered reports and preregistration, these findings also highlight a need for more direct replications. Bias-corrected meta-analysis should generally be viewed as a cautionary note, highlighting the pressing need to revisit findings in a specific area, rather than conclusive evidence in itself. In line with this, we conducted a replication of a study on the Identifiable Victim Effect, in combination with the meta-analysis, and other researchers have taken up the challenge for some of the theories reanalysed here (see for instance Calderon et al., 2023 for a large multi-lab replication project on Construal Level Theory).

7.2.2 Benefits of Computational Modelling

This thesis further highlights the benefits of applying computational modelling to moral and prosocial decisions to understand behaviour on individual level and compare to optimality (in particular in Chapters 5 and 6). In Chapter 5, (dependent) mixture modelling leads to a fine-grained classification of participants’ behaviour in terms of the decision strategies they employ. By formalising the optimal strategies using Bellman Equations and comparing participants’ strategies (as captured by computational models) to them, we can evaluate to what extent participants’ behaviour approximates the optimal solution on an individual level. This led to interesting insights that would have been difficult to discover without this type of modelling (e.g., the higher proportion of participants following an optimal single switch in the Reset than the Keep condition). Chapter 6 highlights one additional benefit of computational models: they allow us to capture individual-level variability in learning style about an intervention and therefore predict for which individuals we would expect transfer to novel situations and for which individuals we would not.

By making precise predictions from computational models, we can make inter-

esting inferences from our data, regardless of how the results turn out. This reduces publication bias and helps cumulative science. In the next section, I outline how computational modelling could be combined with the publication bias adjustment methods outlined in the previous section for an integrative approach that tackles important challenges in the social sciences.

7.3 Where Next? Bridging Publication Bias Adjusted Meta-Analysis and Computational Modelling

The discussion of the empirical and methodological implications from this thesis has highlighted three key issues:

1. **Publication bias.** Results that support a researcher's predictions (usually about the presence of some phenomenon or effect) are much more likely to be published than results that do not support their prediction. This leads to an overestimation of the evidence for and effect size of social science phenomena.
2. **Unexplained variability and unknown moderators.** For many phenomena, researchers do not know under which conditions they can be found or what populations they apply to. This hinders the generalisability and applicability of insights to new intervention types, populations, and domains.
3. **Lack of (computational) theory development.** Theories are powerful tools because they allow researchers to predict what would happen under new conditions that are not actually realised. However, in order to develop good theories, we need (1) a good understanding of the phenomena that the theory aims to explain (including how robust they are in terms of publication bias and their boundary conditions and moderators) and (2) precise mathematical models that make specific predictions (both of which are lacking in many social science disciplines).

It is difficult to tackle each of these problems in isolation, as they amplify each other in various ways: For instance, publication bias makes it more difficult to understand variability in effect sizes and consequently more difficult to develop formal theories to describe this variability. However, the lack of formal theory also amplifies publication bias (as null effects become less informative if the theory is weak) and makes it more difficult to identify moderator variables. To tackle these issues, the main future direction is to integrate meta-analytic work and computational modelling work in research on moral and prosocial decision making and beyond into an integrated methodological pipeline.

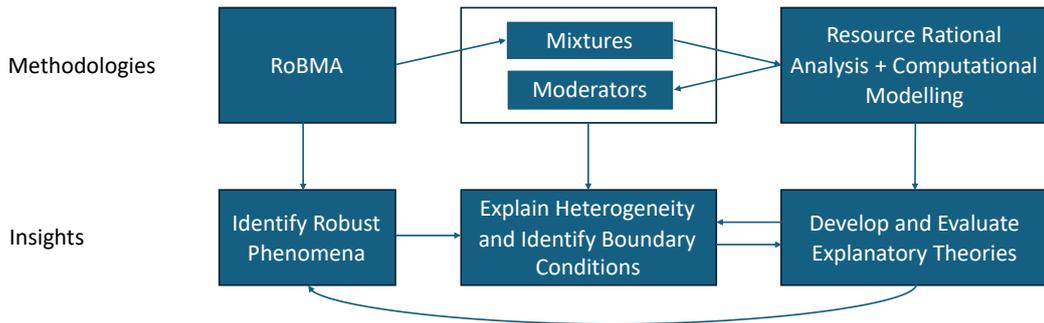
As a first step to doing so, I plan to extend Robust Bayesian Meta-Analysis to include mixture modelling.¹ Given the frequent inadequacy of traditional moderators in explaining heterogeneity among effect sizes (Linden & Hönekopp, 2021; Mertens et al., 2022), mixture modelling offers a data-driven alternative by identifying distinct clusters of effect sizes that can then be interpreted using theory. The insights gained from the mixture models can then be used to build computational models, which in turn allow for new predictions that can be tested using moderator analysis (which we already developed, Bartoš et al., 2025). Integrating mixture meta-analysis and computational modelling will form a cohesive methodological pipeline: the mixture analysis will initially reveal phenomena and moderators, informing subsequent theory development. The resulting computational theories will yield novel, testable predictions that can be validated through further moderator analyses (Figure 7.1).

7.4 Conclusion

This thesis set out to understand why well-intentioned people sometimes devote immense effort and resources to interventions that fail to help (e.g., PlayPump), and conversely, how people and societies can change their moral values (e.g., abolition of slavery). By combining large-scale bias-adjusted meta-analyses, controlled

¹Chapter 3 already included mixture meta-analysis; however, these were frequentist mixtures that did not afford the benefits of Bayesian model-averaging and were more limited in their treatment of publication bias.

Figure 7.1: Where Next? Bridging Meta-Analytic Work and Computational Modelling



behavioural experiments, and formal and computational models, I have taken a deliberately broad approach to understanding moral and prosocial decision making. In Part I, I highlight how several areas of moral and prosocial decision making are affected by publication bias, and I conclude that the evidence base for key theories and interventions is thinner and more context-bound than previously thought. Part II provides a more positive counterpoint: by combining new experimental approaches with computational modelling, I gain new insights and develop interventions in key areas, including scope insensitivity, decision making under extinction risk, and moral learning.

Admittedly, the picture that emerges from Part I can feel bleak: when selective reporting and publication are common, even seminal findings can show little evidence after accounting for publication bias and fail to replicate. However, on the positive side, the field has started to adopt many approaches that can mitigate biases, such as preregistration, direct replications, and registered reports (e.g., Christensen et al., 2020; Nosek et al., 2022). While there is still room for progress, I believe the field is overall moving in the right direction, which is also reflected in the positive reception of the studies presented here, many of which might have faced greater publication barriers a decade ago. Overall, I am therefore optimistic about the development of the field and the future of research on human morality.

The same cautious optimism extends to human morality itself. It has often been argued that human morality is fallible and that people are swayed by morally irrelevant details (Greene, 2013). In contrast, across the new paradigms I intro-

duced, participants often make relatively good decisions. Arguably, many of the paradigms introduced in this thesis are more similar to the types of decisions that people face in the real world (e.g., our experiment on moral learning uses realistic historical dilemmas rather than typical trolley dilemmas), which may be the very reason why people are making better decisions than in much previous work (e.g., Brunswik, 1955; Gigerenzer & Brighton, 2009). This makes me hopeful that, in many real-world contexts, human moral and prosocial decision making is more reasonable than previously thought.

Bibliography

- Aldy, J. E., & Viscusi, W. K. (2007). Age differences in the value of statistical life: Revealed preference evidence. *Review of Environmental Economics and Policy*, 1(241-258). <https://doi.org/10.1093/reep/rem014>
- Alinaghi, N., & Reed, W. R. (2018). Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? *Research Synthesis Methods*, 9(2), 285–311. <https://doi.org/10.1002/jrsm.1298>
- Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1), 73–106. <https://doi.org/10.1162/00335530360535153>
- Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A., Evgeniou, T., Gopnik, A., Jamison, J. C., et al. (2022). Computational ethics. *Trends in Cognitive Sciences*, 26(5), 388–405. <https://doi.org/10.1016/j.tics.2022.02.009>
- AXA Pensions. (2024). Lifecycle strategies [Accessed: 2024-11-04]. <https://pensions.axa-employeebenefits.co.uk/investments/lifecycle-strategies>
- Bakdash, J. Z., & Marusich, L. R. (2022). Left-truncated effects and overestimated meta-analytic means. *Proceedings of the National Academy of Sciences*, 119(31), e2203616119. <https://doi.org/10.1073/pnas.2203616119>
- Banerjee, A., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs.
- Barkan, R., & Busemeyer, J. R. (1999). Changing plans: Dynamic inconsistency and the effect of experience on the reference point. *Psychonomic Bulletin & Review*, 6(4), 547–554. <https://doi.org/10.3758/BF03212962>

- Baron, J., & Greene, J. D. (1996). Determinants of insensitivity to quantity in valuation of public goods: Contribution, warm glow, budget constraints, availability, and prominence. *Journal of Experimental Psychology: Applied*, 2(2), 107. <https://doi.org/10.1037/1076-898X.2.2.107>
- Bartoš, F., Maier, M., Quintana, D., & Wagenmakers, E.-J. (2022). Adjusting for publication bias in JASP and R: Selection models, PET-PEESE, and robust Bayesian meta-analysis. *Advances in Methods and Practices in Psychological Science*, 5(3), 1–19. <https://doi.org/10.1177/25152459221109259>
- Bartoš, F., Maier, M., Shanks, D. R., Stanley, T., Sladekova, M., & Wagenmakers, E.-J. (2023). Meta-analyses in psychology often overestimate evidence for and size of effects. *Royal Society Open Science*, 10(7), 1–12. <https://doi.org/10.1098/rsos.230224>
- Bartoš, F., Maier, M., Stanley, T., & Wagenmakers, E.-J. (2022). Adjusting for publication bias reveals mixed evidence for the impact of cash transfers on subjective well-being and mental health. <https://doi.org/10.31234/osf.io/d9vcg>
- Bartoš, F., Maier, M., Stanley, T., & Wagenmakers, E.-J. (2025). Robust Bayesian meta-regression: Model-averaged moderation analysis in the presence of publication bias. *Psychological Methods*. <https://dx.doi.org/10.1037/met0000737>
- Bartoš, F., Maier, M., Wagenmakers, E.-J., Doucouliagos, H., & Stanley, T. D. (2022). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, 14(1), 99–116. <https://doi.org/10.1002/jrsm.1594>
- Bartoš, F., Maier, M., Wagenmakers, E.-J., Nippold, F., Doucouliagos, H., Ioannidis, J. P. A., Otte, W. M., Sladekova, M., Deressa, T. K., Bruns, S. B., Fanelli, D., & Stanley, T. D. (2024). Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics. *Research Synthesis Methods*, 15, 500–511. <https://doi.org/10.1002/jrsm.1703>

- Bartoš, F., & Maximilian, M. (2020). RoBMA: An R package for robust Bayesian meta-analyses [R package version 2.2.0]. <https://CRAN.R-project.org/package=RoBMA>
- Bartoš, F., & Schimmack, U. (2022). Z-curve. 2.0: Estimating replication rates and discovery rates. *Meta-Psychology*, 6, 1–14. <https://doi.org/10.15626/MP.2021.2720>
- Batson, C. D., & Thompson, E. R. (2001). Why don't moral people act morally? motivational considerations. *Current Directions in Psychological Science*, 10(2), 54–57. <https://doi.org/10.1111/1467-8721.00114>
- Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, 8(9), 536–554. <https://doi.org/10.1111/spc3.12131>
- Bekkers, R. (2017). Words and deeds of generosity: Are decisions about real and hypothetical money really different? [Working paper, Department of Sociology, Utrecht University]. https://renebekkers.files.wordpress.com/2015/12/15_12_01_words_and_deeds.pdf
- Bekkers, R., & Wiepking, P. (2011). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly*, 40(5), 924–973. <https://doi.org/10.1177/0899764010380927>
- Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, 5(2), 187–202. <https://doi.org/10.1177/1745691610362354>
- Bergh, R., & Reinstein, D. (2021). Empathic and numerate giving: The joint effects of victim images and charity evaluations. *Social Psychological and Personality Science*, 12(3), 407–416. <https://doi.org/10.1177/1948550619893968>
- Bernard, M., Dreber, A., Strimling, P., & Eriksson, K. (2013). The subgroup problem: When can binding voting on extractions from a common pool resource

- overcome the tragedy of the commons? *Journal of Economic Behavior & Organization*, *91*, 122–130. <https://doi.org/10.1016/j.jebo.2013.04.009>
- Bhatia, S., & Walasek, L. (2016). Event construal and temporal distance in natural language. *Cognition*, *152*, 1–8. <https://doi.org/10.1016/j.cognition.2016.03.011>
- Blackburn, R. (2010). *The making of new world slavery: From the baroque to the modern, 1492-1800*. Verso Books.
- Blair, R. (2017). Emotion-based learning systems and the development of morality. *Cognition*, *167*, 38–45. <https://doi.org/10.1016/j.cognition.2017.03.013>
- Blais, A.-R., & Weber, E. U. (2006). A domain-specific risk-taking (dospert) scale for adult populations. *Judgment and Decision Making*, *1*(1), 33–47. <https://doi.org/10.1017/S1930297500000334>
- Bland, J. R. (2019). How many games are we playing? an experimental analysis of choice bracketing in games. *Journal of Behavioral and Experimental Economics*, *80*, 80–91. <https://doi.org/10.1016/j.socec.2019.03.011>
- Bodansky, D. (2001). The history of the global climate change regime. In U. Luterbacher & D. F. Sprinz (Eds.), *International relations and global climate change* (pp. 23–41). MIT Press.
- Bom, P. R., & Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*, *10*(4), 497–514. <https://doi.org/10.1002/jrsm.1352>
- Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 1–15. <https://doi.org/10.1057/s41599-021-00903-w>
- Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self-control and meta-decision making. *Trends in Cognitive Sciences*, *19*(11), 700–710. <https://doi.org/10.1016/j.tics.2015.08.013>

- Brighton, H., & Gigerenzer, G. (2012). Homo heuristicus and the bias–variance dilemma. *Action, Perception and the Brain*, 68–91.
- Brodeur, A., Cook, N., Hartley, J., & Heyes, A. (2022). Do pre-registration and pre-analysis plans reduce p-hacking and publication bias? *Available at SSRN*. <http://dx.doi.org/10.2139/ssrn.4180594>
- Brügger, A. (2020). Understanding the psychological distance of climate change: The limitations of construal level theory and suggestions for alternative theoretical perspectives. *Global Environmental Change*, 60, 102023. <https://doi.org/10.1016/j.gloenvcha.2019.102023>
- Brügger, A., Morton, T. A., & Dessai, S. (2016). “proximising” climate change re-considered: A construal level theory perspective. *Journal of Environmental Psychology*, 46, 125–142. <https://doi.org/10.1111/j.1539-6924.2011.01695.x>
- Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology*, 4. <https://doi.org/10.15626/MP.2018.874>
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological review*, 62(3), 193. <https://doi.org/10.1037/h0047470>
- Buchanan, A., & Powell, R. (2018). *The evolution of moral progress: A biocultural theory*. Oxford University Press.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>
- Cai, R., & Leung, X. Y. (2020). Mindset matters in purchasing online food deliveries during the pandemic: The application of construal level and regulatory fo-

- cus theories. *International Journal of Hospitality Management*, 91, 102677. <https://doi.org/10.1016/j.ijhm.2020.102677>
- Calderon, S., Mac Giolla, E., Ask, K., & Granhag, P. A. (2020). Subjective likelihood and the construal level of future events: A replication study of wakslak, trope, liberman, and alony (2006). *Journal of Personality and Social Psychology*, 119(5), e27–e37. <https://doi.org/10.1037/pspa0000214>
- Calderon, S., Mac Giolla, E., Ask, K., & Luke, T. J. (2023). Effects of psychological distance on mental abstraction: A registered report of four tests of construal level theory (stage 1 registered report). <https://doi.org/10.31234/osf.io/wqbhd>
- Callaway, F., Jain, Y. R., van Opheusden, B., Das, P., Iwama, G., Gul, S., Krueger, P. M., Becker, F., Griffiths, T. L., & Lieder, F. (2022). Leveraging artificial intelligence to improve people’s planning strategies. *Proceedings of the National Academy of Sciences*, 119(12), e2117432119.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637–644. <https://doi.org/10.1038/s41562-018-0399-z>
- 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(1), 7–42. <https://doi.org/10.1023/A:1007850605129>
- Cameron, C. D., & Payne, B. K. (2011). Escaping affect: How motivated emotion regulation creates insensitivity to mass suffering. *Journal of Personality and Social Psychology*, 100(1), 7–42. <https://doi.org/10.1023/A:1007850605129>
- Capraro, V., Jagfeld, G., Klein, R., Mul, M., & de Pol, I. v. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific reports*, 9(1), 11880. <https://doi.org/10.1038/s41598-019-48094-4>

- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144. <https://doi.org/10.1177/2515245919847196>
- Caviola, L., Schubert, S., & Nemirow, J. (2020). The many obstacles to effective giving. *Judgment and Decision Making*, 15(2), 159.
- Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49(3), 609–610. <https://doi.org/10.1016/j.cortex.2012.12.016>
- Chambers, C. D. (2017). *The seven deadly sins of psychology*. Princeton University Press.
- Chambers, C. D. (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.
- Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, 66, A1–A2. <https://doi.org/10.1016/j.cortex.2015.03.022>
- Charris, R. A. (2018). *A systematic replication of the identifiable victim effect (small, loewenstein & slovic 2007)* [Doctoral dissertation, University of Los Andes (Colombia)] [Unpublished doctoral dissertation]. <https://repositorio.uniandes.edu.co/bitstream/handle/1992/34740/u808458.pdf?sequence=1>
- Chater, N., & Loewenstein, G. (2022). The i-frame and the s-frame: How focusing on the individual-level solutions has led behavioral public policy astray. *Available at SSRN 4046264*.
- Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97. <https://doi.org/10.1016/j.jbef.2015.12.001>

- Christensen, G., Wang, Z., Levy Paluck, E., Swanson, N., Birke, D., Miguel, E., & Littman, R. (2020). Open science practices are on the rise: The state of social science (3s) survey.
- Christian, B. (2021). *The alignment problem: How can machines learn human values?* Atlantic Books.
- Chu, H., & Yang, J. Z. (2019). Emotion and the psychological distance of climate change. *Science Communication*, *41*(6), 761–789. <https://doi.org/10.1177/1075547019889637>
- Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropogenic shadow: Observation selection effects and human extinction risks. *Risk Analysis*, *30*(10), 1495–1506. <https://doi.org/10.1111/j.1539-6924.2010.01460.x>
- Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037.
- Clark, H., Cárdenas, M., Dybul, M., Kazatchkine, M., Liu, J., Miliband, D., Nordström, A., Sudan, P., Zedillo, E., Obaid, T., et al. (2022). Transforming or tinkering: The world remains unprepared for the next pandemic threat. *The Lancet*, *399*(10340), 1995–1999. [https://doi.org/10.1016/S0140-6736\(22\)00929-1](https://doi.org/10.1016/S0140-6736(22)00929-1)
- Coleman, M., Caviola, L., Lewis, J., & Goodwin, G. (2023). How important is the end of humanity? lay people prioritize extinction prevention but not above all other societal issues. <https://doi.org/10.31234/osf.io/qn7k5>
- Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64.
- Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300.
- Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

- Crockett, M. J. (2016). How formal models can illuminate mechanisms of moral judgment and decision making. *Current directions in psychological science*, 25(2), 85–90. <https://doi.org/10.1177/0963721415624012>
- Crone, D. L., & Laham, S. M. (2017). Utilitarian preferences or action preferences? de-confounding action and moral code in sacrificial dilemmas. *Personality and Individual Differences*, 104, 476–481.
- Crosetto, P., & Filippin, A. (2013). The “bomb” risk elicitation task. *Journal of Risk and Uncertainty*, 47, 31–65. <https://doi.org/10.1007/s11166-013-9170-z>
- Cubitt, R., Starmer, C., & Sugden, R. (1998). Dynamic choice and the common ratio effect: An experimental investigation. *The Economic Journal*, 108(450), 1362–1380. <https://doi.org/10.1111/1468-0297.00346>
- Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, 17(3), 273–292.
- Cushman, F., Kumar, V., & Railton, P. (2017). Moral learning: Psychological and philosophical perspectives. *Cognition*, 167, 1–10. <https://doi.org/10.1016/j.cognition.2017.06.008>
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans’ choices and striatal prediction errors. *Neuron*, 69(6), 1204–1215.
- De Palma, A., Ben-Akiva, M., Brownstone, D., Holt, C., Magnac, T., McFadden, D., Moffatt, P., Picard, N., Train, K., Wakker, P., et al. (2008). Risk, uncertainty and discrete choice models. *Marketing Letters*, 19, 269–285. <https://doi.org/10.1080/10407413.2013.810502>
- Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian q-learning. *Aaai/iaai*, 1998, 761–768.
- DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1), 81–116. <https://doi.org/10.3982/ECTA18709>
- Desvousges, W. H., Johnson, F. R., Dunford, R. W., Hudson, S. P., Wilson, K. N., & Boyle, K. J. (1993). Measuring natural resource damages with contingent

- valuation: Tests of validity and reliability. In B. H. Baltagi & F. Moscone (Eds.), *Contributions to Economic Analysis* (pp. 91–164, Vol. 220). <https://doi.org/10.1016/B978-0-444-81469-2.50009-2>
- Dickert, S., Västfjäll, D., Kleber, J., & Slovic, P. (2012). Valuations of human lives: Normative expectations and psychological mechanisms of (ir) rationality. *Synthese*, *189*(1), 95–105. <https://doi.org/10.1007/s11229-012-0137-4>
- Dickert, S., Västfjäll, D., Kleber, J., & Slovic, P. (2015). Scope insensitivity: The limits of intuitive valuation of human lives in public policy. *Journal of Applied Research in Memory and Cognition*, *4*(3), 248–255. <https://doi.org/10.1016/j.jarmac.2014.09.002>
- Dixit, A. K. (1990). *Optimization in Economic Theory* (2nd edition). Oxford University Press.
- Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.
- Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081. <https://doi.org/10.1016/j.conb.2012.08.003>
- Doucoulagos, C., & Stanley, T. D. (2013). Theory competition and selectivity: Are all economic facts greatly exaggerated? *Journal of Economic Surveys*, *27*, 316–339.
- Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos One*, *18*(3), e0279720.
- Dreber, A., & Nowak, M. A. (2008). Gambling for global goods. *Proceedings of the National Academy of Sciences*, *105*(7), 2261–2262.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. <https://doi.org/10.1111/j.0006-341X.2000.00455.x>
- Eagleton, T. (2008). *The meaning of life: A very short introduction*. OUP Oxford.

- Elga, A., Zhu, J.-Q., & Griffiths, T. L. (2024). People make suboptimal decisions about existential risks. https://www.researchgate.net/publication/384230783_People_Make_Suboptimal_Decisions_about_Existential_Risks
- Elisi, M. (2024). bmsR: Bayesian Model Selection in R [GitHub repository].
- Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4), 643–669. <https://doi.org/10.2307/1884324>
- Ellsberg, D. (2017). *The doomsday machine: Confessions of a nuclear war planner*. Bloomsbury Publishing USA.
- Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, 112(4), 912.
- Erlandsson, A., Björklund, F., & Bäckström, M. (2014). Perceived utility (not sympathy) mediates the proportion dominance effect in helping decisions. *Journal of Behavioral Decision Making*, 27(1), 37–47. <https://doi.org/10.1002/bdm.1789>
- Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 16(4), 779–788. <https://doi.org/10.1177/1745691620970586>
- Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane’s contribution to the Bayes factor hypothesis test. *Statistical Science*, 32, 313–329. <https://doi.org/10.1214/16-STS599>
- Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, 79, 200–216.
- Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1–20.
- Eyal, T., Liberman, N., & Trope, Y. (2008). Judging near and distant virtue and vice. *Journal of Experimental Social Psychology*, 44(4), 1204–1209. <https://dx.doi.org/10.1016/j.jesp.2008.03.012>

- Eyal, T., Liberman, N., & Trope, Y. (2014). Thinking of why a transgression occurred may draw attention to extenuating circumstances: A comment on žeželj & jokić replication. (4), 239–331.
- Falkner, R. (2016). The paris agreement and the new logic of international climate politics. *International Affairs*, 92(5), 1107–1125.
- Fechner, G. T. (1860). *Elemente der psychophysik* (Vol. 2). Breitkopf u. Härtel.
- Feynman, R. P. (1985). "surely you're joking, mr. feynman!": *Adventures of a curious character*. WW Norton & Company.
- Fiedler, K. (2018). How to make psychology a genuine science of behavior: Comment on Dolinski's thoughtful paper. *Social Psychological Bulletin*, 13(2), 1–7. <https://doi.org/10.5964/spb.v13i2.26079>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://doi.org/10.1177/2515245920952393>
- Fogel, R. W. (1994). *Without consent or contract: The rise and fall of american slavery*. WW Norton & Company.
- Foot, P. (1967). The problem of abortion and the doctrine of the double effect.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42.
- Frederick, S., Novemsky, N., Wang, J., Dhar, R., & Nowlis, S. (2009). Opportunity cost neglect. *Journal of Consumer Research*, 36(4), 553–561. <https://doi.org/10.1086/599764>
- Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science advances*, 3(10), e1701381.
- Friedrich, J., & McGuire, A. (2010). Individual differences in reasoning style as a moderator of the identifiable victim effect. *Social Influence*, 5(3), 182–201. <https://doi.org/10.1080/15534511003707352>

- Gainsbury, S. M., Suhonen, N., & Saastamoinen, J. (2014). Chasing losses in online poker and casino games: Characteristics and game play of internet gamblers at risk of disordered gambling. *Psychiatry Research, 217*(3), 220–225. <https://doi.org/10.1016/j.psychres.2014.03.033>
- Galak, J., Small, D., & Stephen, A. T. (2011). Microfinance decision making: A field study of prosocial lending. *Journal of Marketing Research, 48*(SPL), S130–S137. <https://doi.org/10.1509/jmkr.48.SPL.S130>
- Garcia-Marques, L., & Ferreira, M. B. (2018). Is observing behaviour the best way to understand behaviour? *Social Psychological Bulletin, 13*, 1–7. <https://doi.org/10.5964/spb.v13i2.26076>
- Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017a). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology, 113*(3), 343–376. <https://doi.org/10.1037/pspa0000086>
- Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017b). Consequences, norms, and generalized inaction in moral dilemmas: The cni model of moral decision-making. *Journal of Personality and Social Psychology, 113*(3), 343–376.
- Gawronski, B., & Beer, J. S. (2017). What makes moral dilemma judgments “utilitarian” or “deontological”? *Social Neuroscience, 12*(6), 626–632.
- Gawronski, B., Luke, D. M., & Körner, A. (2023). Consequences, norms, and general action tendencies: Understanding individual differences in moral dilemma judgments. In *Motivation and morality: A multidisciplinary approach* (pp. 113–132). American Psychological Association. <https://doi.org/10.1037/0000342-005>
- Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Depart-*

- ment of Statistics, Columbia University, 348.* http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273–278.
- Gibbs, J. C. (2019). *Moral development and reality: Beyond the theories of Kohlberg, Hoffman, and Haidt*. Oxford University Press.
- Gigerenzer, G. (2008). Moral intuition = fast and frugal heuristics? In W. Sinnott-Armstrong (Ed.), *Moral psychology* (pp. 1–26). MIT Press.
- Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*(1), 107–143.
- Gong, H., & Medin, D. L. (2012). Construal levels and moral judgment: Some complications. *Judgment and Decision Making, 7*(5), 628.
- Gong, H., Medin, D. L., Eyal, T., Liberman, N., Trope, Y., Žeželj, I. L., & Jokić, B. R. (2014). Commentaries and rejoinder on žeželj and jokić (2014). *Social Psychology, 45*(4), 327–334. <http://dx.doi.org/10.1027/1864-9335/a000206>
- Greene, J. D. (2002). *The terrible, horrible, no good, very bad truth about morality and what to do about it*. Princeton University.
- Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.
- Grinfeld, G., Wakslak, C., Trope, Y., & Liberman, N. (2021). Construing hypotheticals. *Manuscript Submitted for Publication*. <https://doi.org/10.31234/osf.io/yvafk>
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology, 81*, 80–97. <https://doi.org/10.1016/j.jmp.2017.09.005>
- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridgesampling: An R package for estimating normalizing constants. <https://doi.org/10.48550/arXiv.1710.08162>

- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020a). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10), 1–29. <https://doi.org/10.18637/jss.v092.i10>
- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020b). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10), 1–29. [10.18637/jss.v092.i10](https://doi.org/10.18637/jss.v092.i10)
- Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, 2(1), 123–138. <https://doi.org/10.1080/23743603.2017.1326760>
- Guiso, L., & Paiella, M. (2008). Risk aversion, wealth, and background risk. *Journal of the European Economic Association*, 6(6), 1109–1150. <https://doi.org/10.1162/JEEA.2008.6.6.1109>
- Hallsworth, M. (2023). A manifesto for applying behavioural science. *Nature Human Behaviour*, 7(3), 310–322. <https://doi.org/10.1038/s41562-023-01555-3>
- Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148, 14–31.
- Halpern, D. (2015). *Inside the nudge unit: How small changes can make a big difference*. Random House.
- Harris, J. (1985). *The value of life: An introduction to medical ethics*. Routledge & Kegan Paul.
- Hart, P. S., Lane, D., & Chinn, S. (2018). The elusive power of the individual victim: Failure to find a difference in the effectiveness of charitable appeals focused on one compared to many victims. *PLOS ONE*, 13(7). <https://doi.org/10.1371/journal.pone.0199535>

- Hauser, O. P., Rand, D. G., Peysakhovich, A., & Nowak, M. A. (2014). Cooperating with the future. *Nature*, *511*(7508), 220–223. <https://doi.org/10.1038/nature13530>
- He, R., Jain, Y. R., & Lieder, F. (2021). Measuring and modelling how people learn how to plan and how people adapt their planning strategies to the structure of the environment. *International Conference on Cognitive Modeling*.
- He, R., & Lieder, F. (2023). What are the mechanisms underlying metacognitive learning? *arXiv preprint arXiv:2302.04840*, 45(45).
- Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). Lab.js: A free, open, online study builder. *Behavior Research Methods*, 1–18. <https://doi.org/10.3758/s13428-019-01283-5>
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539. <https://doi.org/10.1111/j.0956-7976.2004.00715.x>
- Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403. <https://doi.org/10.1017/S0140525X01004149>
- Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. <https://doi.org/10.1177/2515245919898657>
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401. <https://doi.org/10.1214/SS%5C%2F1009212519>
- Hong, S., & Reed, W. R. (2020). Using monte carlo experiments to select meta-analytic estimators. *Research Synthesis Methods*, *12*(2), 192–215. <https://doi.org/https://doi.org/10.1002/jrsm.1467>
- Hong, S., & Reed, W. R. (2021). Using monte carlo experiments to select meta-analytic estimators. *Research Synthesis Methods*, *12*(2), 192–215. <https://doi.org/10.1002/jrsm.1467>

- Hotaling, J. M., & Kellen, D. (2022). Dynamic decision making: Empirical and theoretical directions. In *Psychology of learning and motivation* (pp. 207–238, Vol. 76). Elsevier.
- Hsee, C. K., Zhang, J., Lu, Z. Y., & Xu, F. (2013). Unit asking: A method to boost donations and beyond. *Psychological Science*, *24*(9), 1801–1808. <https://doi.org/10.1177/0956797613482947>
- Huh, N., Jo, S., Kim, H., Sul, J. H., & Jung, M. W. (2009). Model-based reinforcement learning under concurrent schedules of reinforcement in rodents. *Learning & Memory*, *16*(5), 315–323.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.
- Imas, A. (2016). The realization effect: Risk-taking after realized versus paper losses. *American Economic Review*, *106*(8), 2086–2109. <https://www.aeaweb.org/articles?id=10.1257/aer.20140386>
- Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*(1), 109–117. <https://doi.org/10.1214/ss/1177013012>
- Jacquet, J., Hagel, K., Hauert, C., Marotzke, J., Röhl, T., & Milinski, M. (2013). Intra-and intergenerational discounting in the climate game. *Nature Climate Change*, *3*(12), 1025–1028. <https://doi.org/10.1038/nclimate2024>
- Jain, Y. R., Gupta, S., Rakesh, V., Dayan, P., Callaway, F., & Lieder, F. (2019). How do people learn how to plan? *Conference on Cognitive Computational Neuroscience*, 826–829.
- Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford University Press.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
- Jenni, K., & Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, *14*, 235–257. <https://doi.org/10.1023/A:1007740225484>
- John, P., Blume, T., et al. (2018). How best to nudge taxpayers? the impact of message simplification and descriptive social norms on payment rates in a cen-

- tral london local authority. *Journal of Behavioral Public Administration*, 1(1). <https://doi.org/10.30636/jbpa.11.10>
- John, T. (2023). Mozi [Accessed: 2024-06-26]. In R. Y. Chappell, D. Meissner, & W. MacAskill (Eds.), *Introduction to utilitarianism*. <https://www.utilitarianism.net/utilitarian-thinker/mozi>
- Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives?
- Jones, C., Hine, D. W., & Marks, A. D. (2017). The future is now: Reducing psychological distance to increase public engagement with climate change. *Risk Analysis*, 37(2), 331–341. <https://doi.org/10.1111/risa.12601>
- Jordan, M. R., Amir, D., & Bloom, P. (2016). Are empathy and concern psychologically distinct? *Emotion*, 16(8), 1107.
- Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social neuroscience*, 10(5), 551–560.
- Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, 125(2), 131–164.
- Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). ‘utilitarian’ judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, 134, 193–209.
- Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American psychologist*, 64(6), 515–526.
- Kahneman, D., & Knetsch, J. L. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management*, 22(1), 57–70. [https://doi.org/10.1016/0095-0696\(92\)90019-S](https://doi.org/10.1016/0095-0696(92)90019-S)
- Kahneman, D., Ritov, I., Schkade, D., Sherman, S. J., & Varian, H. R. (1999). Economic preferences or attitude expressions?: An analysis of dollar responses to public issues. *Journal of Risk and Uncertainty*, (19), 203–242. <https://doi.org/10.1023/A:1007835629236>

- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391.
- Karger, E., Rosenberg, J., Jacobs, Z., Hickman, M., Hadshar, R., Gamin, K., Smith, T., Williams, B., McCaslin, T., & Tetlock, P. (2023). Forecasting existential risks evidence from a long-run forecasting tournament. <https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/64abffe3f024747dd0e38d71/1688993798938/XPT.pdf>
- Karlsson, H., Hellström, S., Moche, H., & Västfjäll, D. (2020). Unit Asking—a method for increasing donations: A replication and extension. *Judgment and Decision Making*, *15*(6), 989–993. <https://doi.org/10.1017/S1930297500008184>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in) variant are subjective representations of described and experienced risk and rewards? *Cognition*, *157*, 126–138. <https://doi.org/10.1016/j.cognition.2016.08.020>
- Keller, A., Marsh, J. E., Richardson, B. H., & Ball, L. J. (2021). A systematic review of the psychological distance of climate change: Towards the development of an evidence-based construct. <https://doi.org/10.31234/osf.io/ntmuy>
- Kemp, L., Xu, C., Depledge, J., Ebi, K. L., Gibbins, G., Kohler, T. A., Rockström, J., Scheffer, M., Schellnhuber, H. J., Steffen, W., et al. (2022). Climate endgame: Exploring catastrophic climate change scenarios. *Proceedings of the National Academy of Sciences*, *119*(34), e2108146119.
- Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a common-sense moral theory. *Cognition*, *167*, 107–123.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings.

- Advances in Methods and Practices in Psychological Science*, 1(4), 443–490. <https://doi.org/10.1177/515245918810225>
- Kloke, J. D., McKean, J. W., et al. (2012). Rfit: Rank-based estimation for linear models. *Journal of R*, 4(2), 57.
- Kniesner, T. J., & Viscusi, W. K. (2019). The value of a statistical life. In *Oxford research encyclopedia of economics and finance*.
- Konovalova, E., & Pachur, T. (2021). The intuitive conceptualization and perception of variance. *Cognition*, 217, 104906. <https://doi.org/10.1016/j.cognition.2021.104906>
- Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS computational biology*, 12(8), e1005090.
- Körner, A., & Deutsch, R. (2023). Deontology and utilitarianism in real life: A set of moral dilemmas based on historic events. *Personality and Social Psychology Bulletin*, 49(10), 1511–1528.
- Krueger, P. M., Callaway, F., Gul, S., Griffiths, T. L., & Lieder, F. (2024). Identifying resource-rational heuristics for risky choice. *Psychological Review*. <https://doi.org/10.1037/rev0000456>
- Krupnick, A. (2007). Mortality-risk valuation and age: Stated preference evidence. *Review of Environmental Economics and Policy*. <https://doi.org/10.1093/reep/rem016>
- Kuhn, E. (1992). *Nietzsches philosophie des europäischen nihilismus [Nietzsche's philosophy of european nihilism]*. De Gruyter.
- Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4), 423–434. <https://doi.org/10.1038/s41562-019-0787-z>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Lange, K., Kühn, S., & Filevich, E. (2015). "just another tool for online studies"(jatos): An easy solution for setup and management of web servers sup-

- porting online studies. *PloS one*, *10*(6), e0130834. <https://doi.org/10.1371/journal.pone.0130834>
- Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ*, *333*(7568), 597–600. <https://doi.org/10.1136/bmj.333.7568.597>
- Leal Filho, W., Sima, M., Sharifi, A., Luetz, J. M., Salvia, A. L., Mifsud, M., Olooto, F. M., Djekic, I., Anholon, R., Rampasso, I., et al. (2021). Handling climate change education at universities: An overview. *Environmental Sciences Europe*, *33*, 1–19. <https://doi.org/10.1186/s12302-021-00552-5>
- Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*(1), 287–308.
- Lee & Feeley, T. H. (2016). The identifiable victim effect: A meta-analytic review. *Social Influence*, *11*(3), 199–215. <https://doi.org/10.1080/15534510.2016.1216891>
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lejarraga, T., & Hertwig, R. (2021). How experimental methods shaped views on human competence and rationality. *Psychological Bulletin*, *147*(6), 535. <https://doi.org/10.1037/bul0000324>
- Lejuez, C. W., Aklin, W. M., Zvolensky, M. J., & Pedulla, C. M. (2003). Evaluation of the Balloon Analogue Risk Task (bart) as a predictor of adolescent real-world risk-taking behaviours. *Journal of Adolescence*, *26*(4), 475–479. [https://doi.org/10.1016/s0140-1971\(03\)00036-8](https://doi.org/10.1016/s0140-1971(03)00036-8)
- Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75. <https://doi.org/10.1037/1076-898x.8.2.75>

- Lesner, T. H., & Rasmussen, O. D. (2014). The identifiable victim effect in charitable giving: Evidence from a natural field experiment. *Applied Economics*, *46*(36), 4409–4430. <https://doi.org/10.1080/00036846.2014.962226>
- Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158–26169.
- Levy, H. (2015). *Stochastic dominance: Investment decision making under uncertainty*. Springer.
- Li, G., Abbade, L. P., Nwosu, I., Jin, Y., Leenus, A., Maaz, M., Wang, M., Bhatt, M., Zielinski, L., Sanger, N., et al. (2018). A systematic review of comparisons between protocols or registrations and full reports in primary biomedical research. *BMC Medical Research Methodology*, *18*(1), 1–20.
- Li, Z., Zhang, S., Liu, X., Kozak, M., & Wen, J. (2020). Seeing the invisible hand: Underlying effects of covid-19 on tourists' behavioral patterns. *Journal of Destination Marketing & Management*, *18*, 100502. <https://doi.org/10.1016/j.jdmm.2020.100502>
- Liberman, N., & Trope, Y. (1998). The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of personality and social psychology*, *75*(1), 5. <https://doi.org/10.1037/0022-3514.75.1.5>
- Liberman, N., & Trope, Y. (2014). Traversing psychological distance. *Trends in cognitive sciences*, *18*(7), 364–369. <https://doi.org/10.1016/j.tics.2014.03.001>
- Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762.
- Lieder, F., Griffiths, T. L., & Hsu, M. (2015). Utility-weighted sampling in decisions from experience. *The 2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making*.

- Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1–32. <https://doi.org/10.1037/rev0000074>
- Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, *14*(4), e1006043.
- Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 281. <https://doi.org/10.1037/cap0000236>
- Lin, Y., Osman, M., & Ashcroft, R. (2017). Nudge: Concept, effectiveness, and ethics. *Basic and Applied Social Psychology*, *39*, 293–306.
- Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, *16*(2), 358–376. <https://doi.org/10.1177/1745691620964193>
- Lister, J. J., Nower, L., & Wohl, M. J. (2016). Gambling goals predict chasing behavior during slot machine play. *Addictive behaviors*, *62*, 129–134. <https://doi.org/10.1016/j.addbeh.2016.06.018>
- Liviatan, I., Trope, Y., & Liberman, N. (2008). Interpersonal similarity as a social distance dimension: Implications for perception of others' actions. *Journal of Experimental Social Psychology*, *44*(5), 1256–1269. <https://doi.org/10.1016/j.jesp.2008.04.007>
- Lockwood, M. (1988). Quality of life and resource allocation. *Royal Institute of Philosophy Supplements*, *23*, 33–55.
- Lockwood, P. L., Apps, M. A., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences*, *113*(35), 9763–9768.
- Lockwood, P. L., Klein-Flügge, M. C., Abdurahman, A., & Crockett, M. J. (2020). Model-free decision making is prioritized when learning to avoid harming

- others. *Proceedings of the National Academy of Sciences*, 117(44), 27719–27730. <https://doi.org/10.1073/pnas.2010890117>
- Lockwood, P. L., van den Bos, W., & Dreher, J.-C. (2024). Moral learning and decision-making across the lifespan. *Annual Review of Psychology*, 76.
- Loewenstein, G., Small, D. A., & Strnad, J. (2006). Statistical, identifiable, and iconic victims. In *Behavioral public finance* (pp. 32–46). <https://dx.doi.org/10.2139/ssrn.678281>
- Lopes, A. F., & Kipperberg, G. (2020). Diagnosing Insensitivity to Scope in Contingent Valuation. *Environmental and Resource Economics*, 77(1), 191–216. <https://doi.org/10.1007/s10640-020-00470-9>
- Loy, L. S., & Spence, A. (2020). Reducing, and bridging, the psychological distance of climate change. *Journal of Environmental Psychology*, 67, 101388. <https://doi.org/10.1016/j.jenvp.2020.101388>
- Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making*, 27(2), 146–156. <https://doi.org/10.1002/bdm.1792>
- Luisetti, T., Bateman, I. J., & Turner, R. K. (2011). Testing the fundamental assumption of choice experiments: Are values absolute or relative? *Land Economics*, 87(2), 284–296.
- MacAskill, W. (2015). *Doing good better: Effective altruism and a radical new way to make a difference*. Guardian Faber Publishing.
- Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2014). Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin & Review*, 21, 629–636. <https://doi.org/10.3758/s13423-013-0542-9>
- Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. *CBE—Life Sciences Education*, 12(3), 345–351. <https://doi.org/10.1187/cbe.13-04-0082>
- Maiella, R., La Malva, P., Marchetti, D., Pomarico, E., Di Crosta, A., Palumbo, R., Cetara, L., Di Domenico, A., & Verrocchio, M. C. (2020). The psy-

- chological distance and climate change: A systematic review on the mitigation and adaptation behaviors. *Frontiers in Psychology*, *11*, 2459. <https://doi.org/10.3389/fpsyg.2020.568899>
- Maier, M., Bartoš, F., Quintana, D. S., Dablander, F., den Bergh, D. v., Marsman, M., Ly, A., & Wagenmakers, E.-J. (2024). Model-averaged bayesian t tests. *Psychonomic Bulletin & Review*, 1–25.
- Maier, M., Bartoš, F., & Wagenmakers, E.-J. (2023). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods*, *28*(1), 107–122. <https://doi.org/10.1037/met0000405>
- Maier, M., Cheung, V., & Lieder, F. (2024). *Metacognitive learning from consequences of past choices shapes moral decision-making* [Conditionally accepted in Nature Human Behaviour]. <https://doi.org/10.31234/osf.io/gjf3h>
- Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2). <https://doi.org/10.31234/osf.io/ts4r6>
- Maier, M., VanderWeele, T. J., & Mathur, M. B. (2022). Using selection models to assess sensitivity to publication bias: A tutorial and call for more routine use. *Campbell Systematic Reviews*, *18*(3), e1256. <https://doi.org/10.1002/cl2.1256>
- Maier, M., Wong, Y. C., & Feldman, G. (2023). Revisiting and rethinking the identifiable victim effect: Replication and extension of small, loewenstein, and slovic (2007). *Collabra: Psychology*, *9*(1). <https://doi.org/10.1525/collabra.90203>
- Marani, M., Katul, G. G., Pan, W. K., & Parolari, A. J. (2021). Intensity and frequency of extreme novel epidemics. *Proceedings of the National Academy of Sciences*, *118*(35), e2105482118. <https://doi.org/10.1073/pnas.2105482118>
- Marcinkiewicz, M. (2016). Increasing charitable donations: The limits of the unit asking effect. *Unpublished Master Thesis*.

- Mata, A., Vaz, A., & Mendonça, B. (2022). Deliberate ignorance in moral dilemmas: Protecting judgment from conflicting information. *Journal of Economic Psychology, 90*, 102523.
- Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 69*(5), 1091–1119.
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.
- McKenzie, C. R., Sher, S., Leong, L. M., & Müller-Trede, J. (2018). Constructed preferences, rationality, and choice architecture. *Review of Behavioral Economics, 5*, 337–360.
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11*(5), 730–749. <https://doi.org/10.1177/17456916166662243>
- Mertens, S., Herberz, M., Hahnel, U. J. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences, 119*(1), e2107346118.
- Meyer, A., & Frederick, S. (2023). The formation and revision of intuitions. *Cognition, 240*, 105380. <https://doi.org/10.1016/j.cognition.2023.105380>
- Michael, O., & Mark, O. (2024). The world is not ready for the next pandemic. *Foreign Affairs*. <https://www.foreignaffairs.com/united-states/world-not-ready-next-pandemic>
- Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A., & Marotzke, J. (2008). The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences, 105*(7), 2291–2294.
- Mill, J. S. (1879). *Utilitarianism*. Fraser's Magazine.

- Millett, P., & Snyder-Beattie, A. (2017a). Existential risk and cost-effective biosecurity. *Health security*, *15*(4), 373–383.
- Millett, P., & Snyder-Beattie, A. (2017b). Human agency and global catastrophic biorisks. *Health security*, *15*(4), 335–336.
- Mitsis, P. (2020). *Oxford handbook of epicurus and epicureanism*. Oxford University Press.
- Moche, H., & Västfjäll, D. (2021). Helping the child or the adult? systematically testing the identifiable victim effect for child and adult victims. *Social Influence*, *16*(1), 78–92. <https://doi.org/10.1080/15534510.2021.1995482>
- Monroe, M. C., Plate, R. R., Oxarart, A., Bowers, A., & Chaves, W. A. (2019). Identifying effective climate change education strategies: A systematic review of the research. *Environmental Education Research*, *25*(6), 791–812. <https://doi.org/10.1080/13504622.2017.1360842>
- Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, *16*(4), 707–716. <https://doi.org/10.1177/174569162097476>
- Nichols, S. (2021). *Rational rules: Towards a theory of moral learning*. Oxford University Press.
- Nichols, S. (2022). Moral learning and moral representations. *The Oxford Handbook of Moral Psychology*, 421.
- Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. <https://doi.org/10.1016/j.tics.2019.07.009>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The pre-registration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M.,

- Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- O’Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, *1*, 94–100. <https://doi.org/10.1016/j.cobeha.2014.10.004>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). <https://doi.org/10.1126/science.1205596>
- Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173–182. <https://doi.org/10.1038/s41562-018-0506-1>
- Ord, T. (2009). Beyond action: Applying consequentialism to decision making and motivation.
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. <https://doi.org/10.1177/1745691612463401>
- Patil, I., Zucchelli, M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2020). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology*, *120*. <https://doi.org/10.1037/pspp0000281>
- Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*(11), 803–809. <https://doi.org/10.1038/s41562-017-0219-x>
- Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., & Leff, A. P. (2010). Comparing families of dynamic causal models. *PLoS Computational Biology*, *6*(3), e1000709.

- Peress, J. (2004). Wealth, information acquisition, and portfolio choice. *The Review of Financial Studies*, 17(3), 879–914. <https://doi.org/10.1093/rfs/hhg056>
- Perfors, A., & Van Dam, N. T. (2018). Human decision making in black swan situations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 40. <https://escholarship.org/uc/item/3p76280v>
- Perry, W. J., & Collina, T. Z. (2020). *The Button: The New Nuclear Arms Race and Presidential Power from Truman to Trump*. BenBella Books.
- Persson, E., & Tinghög, G. (2020). Opportunity cost neglect in public policy. *Journal of Economic Behavior & Organization*, 170, 301–312. <https://doi.org/10.1016/j.jebo.2019.12.012>
- Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 167. <https://doi.org/10.1037/0278-7393.34.1.167>
- Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and Clinical Psychopharmacology*, 16(6), 555. <https://doi.org/10.1037/a0014245>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Rabin, M., & Weizsäcker, G. (2009). Narrow bracketing and dominated choices. *American Economic Review*, 99(4), 1508–1543. <https://doi.org/10.1257/aer.99.4.1508>
- Raftery, A. E., Madigan, D., & Volinsky, C. T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. *In Bayesian Statistics 5*, 323–349.
- Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. *Cognition*, 167, 172–190.

- Read, D., Loewenstein, G., Rabin, M., Keren, G., & Laibson, D. (2000). Choice bracketing. In B. Fischhoff & C. F. Manski (Eds.), *Elicitation of preferences* (pp. 171–202). Springer. https://doi.org/10.1007/978-94-017-1406-8_7
- Reczek, R. W., Trudel, R., & White, K. (2018). Focusing on the forest or the trees: How abstract versus concrete construal level predicts responses to eco-friendly products. *Journal of Environmental Psychology, 57*, 87–98. <https://doi.org/10.1016/j.jenvp.2018.06.003>
- Rickard, L. N., Yang, Z. J., & Schuldt, J. P. (2016). Here and now, there and then: How “departure dates” influence climate change engagement. *Global Environmental Change, 38*, 97–107. <https://doi.org/10.1016/j.gloenvcha.2016.03.003>
- Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1446. <https://doi.org/10.1037/a0013646>
- Rieskamp, J., & Otto, P. E. (2006). Ssl: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*(2), 207.
- Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage, 84*, 971–985.
- Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research, 21*(4), 37–59. https://doi.org/10.1300/J079v21n04_02
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of levels of significance. *Psychological Reports, 15*(2), 570. <https://doi.org/10.2466/pr0.1964.15.2.570>
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. <https://doi.org/10.3758/s13423-014-0595-4>

- Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, *73*(2), 186–190. <https://doi.org/10.1080/00031305.2017.1341334>
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.
- Sánchez, A. M., Coleman, C. W., & Ledgerwood, A. (2021). Does temporal distance influence abstraction? a large pre-registered experiment. *Social Cognition*, *39*(3), 352–365. <https://doi.org/10.1521/soco.2021.39.3.352>
- Sartre, J.-P. (1972). *Being and nothingness*. Pocket Books.
- Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.
- Schinkel, A., & de Ruyter, D. J. (2017). Individual moral development and moral progress. *Ethical Theory and Moral Practice*, *20*, 121–136.
- Schoenegger, P., & Costa-Gomes, M. (2022). Sure-thing vs. probabilistic charitable giving: Experimental evidence on the role of individual differences in risky and ambiguous charitable decision-making. *PloS One*, *17*(9), e0273971. <https://doi.org/10.1371/journal.pone.0273971>
- Schuldt, J. P., Rickard, L. N., & Yang, Z. J. (2018). Does reduced psychological distance increase climate engagement? on the limits of localizing climate change. *Journal of Environmental Psychology*, *55*, 147–153. <https://doi.org/10.1016/j.jenvp.2018.02.001>
- Schultner, D. T., Lindström, B. R., Cikara, M., & Amodio, D. M. (2024). Transmission of social bias through observational learning. *Science Advances*, *10*(26), eadk2030.
- Schulze, C., van Ravenzwaaij, D., & Newell, B. R. (2015). Of matchers and maximizers: How competition shapes choice under risk and uncertainty. *Cognitive Psychology*, *78*, 78–98. <https://doi.org/10.1016/j.cogpsych.2015.03.002>
- Seneca, L. A., et al. (2004). *On the shortness of life*. Penguin UK.
- Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*, 99–124.

- Shipley, N. J., & van Riper, C. J. (2022). Pride and guilt predict pro-environmental behavior: A meta-analysis of correlational and experimental evidence. *Journal of Environmental Psychology, 79*, 101753. <https://doi.org/10.1016/j.jenvp.2021.101753>
- Simmons, J. P. (2013). A New Way To Increase Charitable Donations: Does It Replicate? <http://datacolada.org/2013/10/02/3-a-new-way-to-increase-charitable-donations-does-it-replicate/>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Singer, P. (2019). *The life you can save: How to do your part to end world poverty*. The Life You Can Save.org.
- Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). *Afex: Analysis of factorial experiments* [R package version 1.1-1]. <https://CRAN.R-project.org/package=afex>
- Singmann, H., Xiong, Y., Song, Y., Breen, M., & Baumann, C. (2024). Full-information optimal-stopping problems: Providing people with the optimal policy does not improve performance. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46*. <https://escholarship.org/uc/item/6cs782x5>
- Skinner, B. F. (1963). Operant behavior. *American psychologist, 18*(8), 503.
- Skinner, B. F. (1957). The experimental analysis of behavior. *American Scientist, 45*(4), 343–371.
- Slovic, P. (2007). If i look at the mass i will never act: Psychic numbing and genocide. *Judgment and Decision Making, 2*(2), 1–17. <https://journal.sjdm.org/7303a/jdm7303a.html>
- Slovic, P., & Västfjäll, D. (2010a). Affect, moral intuition, and risk. *Psychological Inquiry, 21*(4), 387–398. <https://doi.org/10.1080/1047840X.2010.521119>

- Slovic, P., & Västfjäll, D. (2010b). Affect, moral intuition, and risk. *Psychological Inquiry*, 21(4), 387–398. <https://doi.org/10.1080/1047840X.2010.521119>
- Slovic, P., & Weber, E. U. (2013). Perception of risk posed by extreme events. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2293086
- Small, D. A., & Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, 26, 5–16. <https://doi.org/10.1023/A:1022299422219>
- Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, 102(2), 143–153. <https://doi.org/10.1016/j.obhdp.2006.01.005>
- Smart, J. J. C., & Williams, B. (1973). *Utilitarianism: For and against*. Cambridge University Press.
- Snefjella, B., & Kuperman, V. (2015). Concreteness and psychological distance in natural language use. *Psychological Science*, 26(9), 1449–1460. <https://doi.org/10.1177/0956797615591771>
- Soderberg, C. K., Callahan, S. P., Kochersberger, A. O., Amit, E., & Ledgerwood, A. (2015). The effects of psychological distance on abstraction: Two meta-analyses. *Psychological bulletin*, 141(3), 525.
- Spence, A., Poortinga, W., & Pidgeon, N. (2012). The psychological distance of climate change. *Risk Analysis*, 32(6), 957–972.
- Spiller, S. A. (2019). Opportunity cost neglect and consideration in the domain of time. *Current Opinion in Psychology*, 26, 98–102. <https://doi.org/10.1016/j.copsy.2018.10.001>
- Stagnaro, M. N., Druckman, J., Berinsky, A. J., Arechar, A. A., Willer, R., & Rand, D. (2024). Representativeness versus attentiveness: A comparison across nine online survey samples. <https://doi.org/10.31234/osf.io/h9j2d>
- Stan Development Team. (2024). RStan: The R interface to Stan [R package version 2.32.6]. <https://mc-stan.org/>

- Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8(5), 581–591. <https://doi.org/10.1177/1948550617693062>
- Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, 144(12), 1325–1346. <https://doi.org/10.1037/bul0000169>
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60–78. <https://doi.org/10.1002/jrsm.1095>
- Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, 36(10), 1580–1598. <https://doi.org/10.1002/sim.7228>
- Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2022). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, 27(2), 177–197. <https://doi.org/10.1037/met0000354>
- Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of *p*-hacking strategies. *Royal Society Open Science*, 10(2), 1–30. <https://doi.org/10.1098/rsos.220346>
- Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton university press.
- Sudhir, K., Roy, S., & Cherian, M. (2016). Do sympathy biases induce charitable giving? the effects of advertising content. *Marketing Science*, 35(6), 849–869. <https://doi.org/10.1287/mksc.2016.0989>
- Sundh, J. (2024). Human behavior in the context of low-probability high-impact events. *Humanities and Social Sciences Communications*, 11(1), 1–10. <https://doi.org/10.1057/s41599-024-03403-9>
- Sunstein, C. R., & Zeckhauser, R. (2011). Overreaction to fearsome risks. *Environmental and Resource Economics*, 48, 435–449. <https://doi.org/10.1007/s10640-010-9449-3>

- Sussman, A. B., Sharma, E., & Alter, A. L. (2015). Framing charitable donations as exceptional expenses increases giving. *Journal of Experimental Psychology: Applied*, 21(2), 130. <https://doi.org/10.1037/xap0000047>
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT press.
- Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., & Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, 119(31), e2200732119.
- Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology*, 58(9), 894–901. <https://doi.org/10.1016/j.jclinepi.2005.01.006>
- Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.
- Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing science*, 4(3), 199–214. <https://doi.org/10.1287/mksc.4.3.199>
- Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, 12(3), 183–206.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- Thomas, K. J., Hamilton, B. C., & Loughran, T. A. (2018). Testing the transitivity of reported risk perceptions: Evidence of coherent arbitrariness. *Criminology*, 56(1), 59–86. <https://doi.org/10.1111/1745-9125.12154>
- Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 204–217.
- Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. OUP USA.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208.

- Trope, Y., & Liberman, N. (2003). Construal level theory of intertemporal judgment and decision. In G. Loewenstein, D. Read, & R. Baumeister (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice* (pp. 245–276, Vol. 1).
- Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological review*, *117*(2), 440. <https://dx.doi.org/10.1037/a0018963>
- Trope, Y., & Liberman, N. (2011). Construal level theory. In P. Van Lange, A. W. Kruglanski, & E. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 118–134, Vol. 1). <https://dx.doi.org/10.4135/9781446249215.n7>
- Tusche, A., & Bas, L. M. (2021). Neurocomputational models of altruistic decision-making and social motives: Advances, pitfalls, and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(6), e1571. <https://doi.org/10.1002/wcs.1571>
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. <https://doi.org/10.1007/BF00122574>
- UNICEF. (2007, October). *An evaluation of the playpump® water system as an appropriate technology for water, sanitation and hygiene programmes* (tech. rep.) (Accessed: 2025-05-19). United Nations Children’s Fund (UNICEF). New York, NY. https://www-tc.pbs.org/frontlineworld/stories/southernafrica904/flash/pdf/unicef_pp_report.pdf
- van den Akker, O. R., Bakker, M., van Assen, M. A., Pennington, C. R., Verweij, L., Elsherif, M. M., Claesen, A., Gaillard, S. D., Yeung, S. K., Frankenberger, J.-L., et al. (2024). The potential of preregistration in psychology: Assessing preregistration producibility and preregistration-study consistency. *Psychological Methods*. <https://doi.org/10.1037/met0000687>
- van den Akker, O. R., van Assen, M. A., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2024). Preregistration in practice: A comparison of pre-registered and non-preregistered studies in psychology. *Behavior Research Methods*, *56*(6), 5424–5433. <https://doi.org/10.3758/s13428-023-02277-0>

- van den Akker, O. R., van Assen, M. A., Enting, M., de Jonge, M., Ong, H. H., Ruffer, F., Schoenmakers, M., Stoevenbelt, A. H., Wicherts, J. M., & Bakker, M. (2023). Selective hypothesis reporting in psychology: Comparing preregistrations and corresponding publications. *Advances in Methods and Practices in Psychological Science*, 6(3), 25152459231187988. <https://doi.org/10.1177/25152459231187988>
- Van Lange, P. A., & Huckelba, A. L. (2021). Psychological distance: How to make climate change less abstract and closer to the self. *Current Opinion in Psychology*, 42, 49–53. <https://doi.org/10.1016/j.copsyc.2021.03.011>
- Van Lent, L. G., Sungur, H., Kunneman, F. A., Van De Velde, B., & Das, E. (2017). Too far to care? Measuring public attention and fear for ebola using twitter. *Journal of Medical Internet Research*, 19(6), e7219. <https://doi.org/10.2196/jmir.7219>
- Vanguard. (2021). Vanguard's life-cycle investing model (vlcm): A general portfolio framework [Accessed: 2024-11-04]. https://corporate.vanguard.com/content/dam/corp/research/pdf/Vanguards-Life-Cycle-Investing-Model-VLCM-A-general-portfolio-framework-US-ISGVLCM_032021_online.pdf
- Västfjäll, D., & Slovic, P. (2020). A psychological perspective on charitable giving and monetary donations: The role of affect. In T. Zaleskiewicz & J. Traczyk (Eds.), *Psychological perspectives on financial decision making* (pp. 331–345). Springer. https://doi.org/10.1007/978-3-030-45500-2_14
- Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13(4), 411–417. <https://doi.org/10.1177/1745691617751884>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistical Computing*, 27, 1413–1432.
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved \hat{R} for assessing con-

- vergence of mcmc. *Bayesian analysis*, 1(1), 1–28. <https://doi.org/10.1214/20-BA1221>
- Veisten, K., Hoen, H. F., Navrud, S., & Strand, J. (2004). Scope insensitivity in contingent valuation of complex environmental amenities. *Journal of Environmental Management*, 73(4), 317–331. <https://doi.org/10.1016/j.jenvman.2004.07.008>
- Verbeke, P., & Verguts, T. (2024). Reinforcement learning and meta-decision-making. *Current Opinion in Behavioral Sciences*, 57, 101374. <https://doi.org/10.1016/j.cobeha.2024.101374>
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419–435. <https://doi.org/10.1007/BF02294384>
- Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, 10(4), 428. <https://psycnet.apa.org/doi/10.1037/1082-989X.10.4.428>
- Wagenmakers, E.-J. (2020). Bayesian thinking for toddlers.
- Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25(3), 169–176. <https://doi.org/10.1177/0963721416643289>
- Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, 605(7910), 423–425. <https://doi.org/10.1038/d41586-022-01332-8>
- Wagenmakers, E.-J., van Ravenzwaaij, D., & Ron, J. (2020, May). Concerns about the default cauchy are often exaggerated: A demonstration with jasp 0.12 [Bayesian Spectacles]. <https://www.bayesianspectacles.org/concerns-about-the-default-cauchy-are-often-exaggerated-a-demonstration-with-jasp-0-12/>
- Wakslak, C. J., Trope, Y., Liberman, N., & Alony, R. (2006). Seeing the forest when entry is unlikely: Probability and the mental representation of events.

- Journal of Experimental Psychology: General*, 135(4), 641. <https://doi.org/10.1037/0096-3445.135.4.641>
- Walker, J. M., Gardner, R., Herr, A., & Ostrom, E. (2000). Collective choice in the commons: Experimental results on proposed allocation rules and votes. *The Economic Journal*, 110(460), 212–234.
- Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review*, 112(4), 862. <https://psycnet.apa.org/doi/10.1037/0033-295X.112.4.862>
- Wang, S., Hurlstone, M. J., Leviston, Z., Walker, I., & Lawrence, C. (2019). Climate change from a distance: An analysis of construal level and psychological distance from climate change. *Frontiers in Psychology*, 10, 230. <https://doi.org/10.3389/fpsyg.2019.00230>
- Wang-Ly, N., & Newell, B. R. (2024). Income volatility and saving decisions: Experimental evidence. *Journal of Behavioral and Experimental Finance*, 43, 100941. <https://doi.org/10.1016/j.jbef.2024.100941>
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44, 92–107. <https://doi.org/10.1006/jmps.1999.1278>
- Watkins, C. J. C. H. (1989). Learning from delayed rewards.
- Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine learning*, 8, 279–292. <https://doi.org/10.1007/BF00992698>
- Weber, E. U. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change*, 77(1), 103–120. <https://doi.org/10.1007/s10584-006-9060-3>
- Weber, E. U. (2010). What shapes perceptions of climate change? *Wiley Interdisciplinary Reviews: Climate Change*, 1(3), 332–342.
- Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4), 263–290. <https://doi.org/10.1002/bdm.414>

- Weiss, D. J., & Shanteau, J. (2021). The futility of decision making research. *Studies in History and Philosophy of Science Part A*, 90, 10–14. <https://doi.org/https://doi.org/10.1016/j.shpsa.2021.08.018>
- Wiener, J. B. (2016). The tragedy of the uncommons: On the politics of apocalypse. *Global Policy*, 7, 67–80.
- Williams, A. (1997). Intergenerational equity: An exploration of the ‘fair innings’ argument. *Health economics*, 6(2), 117–132.
- Williams, E. G. (2023). Rule utilitarianism and rational acceptance. *The Journal of Ethics*, 1–24. <https://doi.org/10.1007/s10892-023-09428-7>
- Wilson, P. A. (2023). The anthropic principle. In *Cosmology* (pp. 505–514). CRC Press.
- Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42, 369–390. <https://doi.org/10.1080/14786442108633773>
- Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, 144(2), 140–176. <https://doi.org/10.1037/bul0000115>
- Yang, J. Z., Rickard, L. N., Liu, Z., & Boze, T. (2021). Too close to care? A replication study to re-examine the effect of cued distance on climate change engagement. *Environmental Communication*, 15(1), 1–11. <https://doi.org/10.1080/17524032.2020.1777181>
- Yeomans, M. (2021). A concrete example of construct construction in natural language. *Organizational Behavior and Human Decision Processes*, 162, 81–94. <https://doi.org/10.1016/j.obhdp.2020.10.008>
- Žeželj, I. L., & Jokić, B. R. (2014). Replication of experiments evaluating impact of psychological distance on moral judgment. *Social Psychology*, 45(3), 223–231. <http://dx.doi.org/10.1027/1864-9335/a000188>

Appendix A

Construal Level Theory

A.1 Additional RoBMA Results

Table A.1: Summary of RoBMA Estimates for Soderberg et al. (2015).

Abstraction	All	Most Precise	Aggregated
BF ₀₁	7.75	9.09	6.57
BF _{rf}	$> 10^{100}$	3.70×10^6	$> 10^{100}$
BF _{pb}	$> 10^{100}$	22826	4.19×10^6
δ (95% CI)	-0.002 [-0.113, 0.056]	0.004 [-0.036, 0.116]	0.003 [-0.093, 0.149]
τ (95% CI)	0.367 [0.314, 0.421]	0.217 [0.142, 0.347]	0.437 [0.251, 0.557]

Downstream Consequences	All	Most Precise	Aggregated
BF ₀₁	0.061	5.92	4.03
BF _{rf}	$> 10^{100}$	$> 10^{100}$	$> 10^{100}$
BF _{pb}	$> 10^{100}$	4.34×10^6	$> 10^{100}$
δ (95% CI)	-0.360 [-0.657, 0.000]	0.012 [-0.004, 0.167]	-0.021 [-0.054, 0.247]
τ (95% CI)	0.712 [0.616, 0.804]	0.311 [0.240, 0.387]	0.398 [0.316, 0.461]

Note. δ and τ correspond to mean model-averaged estimates with 95% credible intervals. BF₀₁ corresponds to the Bayes factor *against* an effect, BF_{rf} corresponds to the Bayes factor for heterogeneity, and BF_{pb} to the Bayes factor for publication bias. The Bayes factors equal to $> 10^{100}$ correspond to values beyond the numerical precision of \mathbb{R} .

A.2 Coding Procedure RoBMA

We replicated the meta-analytic search using the web of science search terms in the appendix of Soderberg et al. (2015). Unfortunately, some of the search terms provided by Soderberg included typos. We contacted the authors of the original paper for clarification but did not receive a reply. Therefore, we fixed these typos ourselves and proceeded with the following topic search string:

```
('`construal level theory`' OR '`construal level`') OR
```

Table A.2: Summary of RoBMA Estimates for Soderberg et al. (2015) with $|\delta| < 1.5$.

Abstraction	All	Most Precise	Aggregated
BF ₀₁	0.240	5.29	14.71
BF _{rf}	$> 10^{100}$	35293	$> 10^{100}$
BF _{pb}	1.40×10^6	24946	$> 10^{100}$
δ (95% CI)	0.143 [0.000, 0.279]	0.020 [0.000, 0.237]	0.003 [0.000, 0.053]
τ (95% CI)	0.232 [0.166, 0.302]	0.155 [0.091, 0.265]	0.158 [0.116, 0.205]

Downstream Consequences	All	Most Precise	Aggregated
BF ₀₁	1.17	0.312	0.032
BF _{rf}	$> 10^{100}$	740002	$> 10^{100}$
BF _{pb}	$> 10^{100}$	26900	$> 10^{100}$
δ (95% CI)	0.068 [0.000, 0.238]	0.153 [0.000, 0.312]	0.276 [0.000, 0.373]
τ (95% CI)	0.315 [0.251, 0.372]	0.212 [0.136, 0.304]	0.223 [0.171, 0.317]

Note. δ and τ correspond to mean model-averaged estimates with 95% credible intervals. BF₀₁ corresponds to the Bayes factor *against* an effect, BF_{rf} corresponds to the Bayes factor for heterogeneity, and BF_{pb} to the Bayes factor for publication bias. The Bayes factors equal to $> 10^{100}$ correspond to values beyond the numerical precision of R.

Table A.3: Summary of RoBMA Estimates for Soderberg et al. (2015) When Using Only Selection Models.

Abstraction	All	Most Precise	Aggregated
BF ₀₁	0.292	2.54	6.25
BF _{rf}	$> 10^{100}$	$> 10^{100}$	$> 10^{100}$
BF _{pb}	1099	1596	6816563
δ (95% CI)	0.175 [0.000, 0.365]	0.046 [0.000, 0.295]	0.002 [-0.118, 0.153]
τ (95% CI)	0.490 [0.409, 0.583]	0.373 [0.264, 0.474]	0.472 [0.393, 0.562]

Downstream Consequences	All	Most Precise	Aggregated
BF ₀₁	0.061	5.93	4.04
BF _{rf}	$> 10^{100}$	$> 10^{100}$	$> 10^{100}$
BF _{pb}	$> 10^{100}$	8678036	$> 10^{100}$
δ (95% CI)	-0.360 [-0.657, 0.000]	0.012 [-0.004, 0.167]	-0.021 [-0.054, 0.247]
τ (95% CI)	0.712 [0.616, 0.804]	0.311 [0.240, 0.385]	0.398 [0.316, 0.461]

Note. δ and τ correspond to mean model-averaged estimates with 95% credible intervals. BF₀₁ corresponds to the Bayes factor *against* an effect, BF_{rf} corresponds to the Bayes factor for heterogeneity, and BF_{pb} to the Bayes factor for publication bias. The Bayes factors equal to $> 10^{100}$ correspond to values beyond the numerical precision of R.

(``construal level theory'' OR ``construal level'' AND abstract) OR (``psychological distance'' AND ``abstract'' AND (``Spatial'' OR ``social'' OR ``temporal'' OR ``probability'' OR ``power'')) OR (``psychological distance'' AND ``concrete'' AND (``Spatial'' OR ``social'' OR ``temporal'' OR ``probability'' OR ``power'')) OR (``Gestalt completion task'' OR ``BIF'' OR ``Vallacher and Wegner, 1989'' OR ``Navon Task'' AND

Table A.4: Summary of RoBMA Estimates for Soderberg et al. (2015) When Using Only Selection Models with $|\delta| < 1.5$.

Abstraction	All	Most Precise	Aggregated
BF ₀₁	0.161	0.754	1.40
BF _{rf}	$> 10^{100}$	27441367	$> 10^{100}$
BF _{pb}	2588362	7986	6581221
δ (95% CI)	0.154 [0.000, 0.279]	0.106 [0.000, 0.290]	0.072 [0.000, 0.282]
τ (95% CI)	0.233 [0.168, 0.304]	0.217 [0.129, 0.316]	0.284 [0.198, 0.369]

Downstream Consequences	All	Most Precise	Aggregated
BF ₀₁	1.17	0.288	0.032
BF _{rf}	$> 10^{100}$	$> 10^{100}$	$> 10^{100}$
BF _{pb}	$> 10^{100}$	52154	$> 10^{100}$
δ (95% CI)	0.068 [0.000, 0.238]	0.157 [0.000, 0.312]	0.276 [0.000, 0.373]
τ (95% CI)	0.315 [0.251, 0.372]	0.214 [0.142, 0.304]	0.223 [0.171, 0.317]

Note. δ and τ correspond to mean model-averaged estimates with 95% credible intervals. BF₀₁ corresponds to the Bayes factor *against* an effect, BF_{rf} corresponds to the Bayes factor for heterogeneity, and BF_{pb} to the Bayes factor for publication bias. The Bayes factors equal to $> 10^{100}$ correspond to values beyond the numerical precision of R.

```
((`construal` OR `abstract` OR `concrete`)) OR ((`hidden
figures` OR `embedded figures` OR `gestalt completion`
OR `snowy pictures` OR `Analysis-Holism Scale`) AND
`construal`) OR (`Levels of personal agency: Individual
variation in action identification`)
```

MO then checked the papers against the inclusion criteria of Soderberg et al. (2015, p.531). Finally, MM double checked effect sizes at random. MM checked sequentially comparing the alternative hypothesis of an accuracy of .9 to uniform prior over accuracy until Bayes factor 10 for (or against) sufficient accuracy was reached.

A.3 Coding Procedure Z-curve

We used search for the topic (`construal level theory` or `construal level`) on Web of Science. Then we sorted the resulting articles based on citations/year and read the abstracts in descending order. We identified 400 relevant articles based on their abstracts. MM & MO extracted the first test statistic from each article that related to construal level as dependent variable, independent variable, or moderator (306 out of the 400 articles). Often no test

statistic could be extracted, mostly due to inadequate reporting (e.g., only $p < .05$).

Appendix B

Sequential Unit Asking

B.1 Pilot 2 - Budget Constrains as Explanation for The Effectiveness of SUA over CUA

In pilot two, we aimed to investigate whether a possible mechanism explaining the effectiveness of SUA was that people disengage when they need to increase the donation too much in one step. In other words, they might donate beyond budget constraints. We manipulated this by starting out from a small donation of \$0.1 for one child resulting in a small donation even after multiplying with 100 versus a larger donation of \$10 where people should disengage when directly jumping to a consistent donation of \$1000 for the full scope.

B.1.1 Methods

Experimental Conditions

This study employed a 2 (technique: SUA vs. CUA) x 2 (costs: small vs. large) between-subjects design. We manipulated the costs required to donate to help one child. The vignette was similar to Study 1, with the modification that this time participants were asked to donate to help treat parasitic worms, with treatment costing either large (\$10) or small (\$0.1). One group was told that 10cts are needed to help one child that is sick of parasitic worms, while the other group that \$10 are needed to cure this disease. In addition, we varied SUA vs. CUA. We hypothesized that both interventions should work similarly when only 10cts were required to help one

child but that SUA would work remarkably better when \$10 was required to help one child. In other words, there would be an interaction between Asking Type (SUA vs. CUA) and the costs of saving one child. We also randomized whether money was displayed in cents or in dollars to check for an effect of the currency unit. In addition, for our analysis, we rescaled the group giving only \$0.1 by multiplying it by 1000 so that the interaction effect could be analyzed.

The vignette was similar to Study 1, with the modification that this time participants were asked to donate to help treat parasitic worms, with treatment costing either large (\$10) or small (\$0.1). In addition, all values were presented on the cents scale and the dollar scale since we had suspected that participants are primed to donate more when the currency unit is dollar compared to cents. To check this idea, half of the participants saw all units in dollars and the other half in cents.

Analyses

We used a Bayesian ANOVA in the package `BRMS` as in the previous study. Because we cannot meaningfully analyze scope insensitivity for participants that donated 0 (as they would donate 0 for any scope under scope consistency as well as scope inconsistency), we excluded these participants. We excluded unbelievably large WTD judgements of more than \$10,000. As our data was (after excluding the zeros) expected to be positive and quite skewed, we used a lognormal likelihood rather than the more common Gaussian likelihood. We used a prior of normal(3,1) on the intercept, normal(0, 0.4) on the scope (as we expected scope insensitivity) and normal(0.4, 0.4) on the main effect of Asking Type, normal(0.25, 0.25) on the interaction and normal(1, 0.5) on σ . The reason why the priors are increasingly more narrow is that given the lognormal likelihood the prediction on the linear scale corresponds to the exponent of the marginal mean. Therefore, using similar priors on main effects in interactions would result in the prediction of an extremely large interaction effect on the linear scale. We tested for the presence of the hypothesized interactions by comparing models that include the interaction to models that do not include the interaction using Bayes factors. We also conducted a non-parametric

analysis using robust ANOVA in the `Rfit` package.

Participants

The study was run on the 23rd of January 2021, targeting American participants through posity. We paid participants \$0.31 to participate in a 2.5-minute study. 2 participants were excluded because they indicated extremely high or negative age. The mean participant age was 443.12 ($sd = 14.14$). 97 participants were female, 103 were male, and 1 did not indicate their gender. This study was approved by University of Oxford Central University Research Ethics Committee (reference number R56657/RE001).

B.1.2 Results and Discussion

A Bayesian contingency table analysis indicated that the intervention did not influence the share of participants donating ($BF_{10} = 0.19$). Therefore, we proceeded to the main analysis, excluding 25 participants that donated 0 and 8 participants that donated more than 10,000. We find no evidence for a main effect of SUA vs. CUA ($BF_{10} = 1.11$) but evidence that people donate more when currencies are presented in terms of dollars rather than cents ($BF_{10} = 8.67$). We find weak evidence against the predicted interaction of Asking Type with scope ($BF_{10} = 0.55$). This is also corroborated by the robust ANOVA which finds no evidence for a main effect of Asking Type, $F(1, 163) = 0.22$, $p = .637$, a main effect of cents vs. dollars, $F(1, 163) = 4.02$, $p = .047$, and no evidence for an interaction effect of starting value and Asking Type, $F(1, 163) = 0.010$, $p = .752$.

To sum up, this study neither found a main effect of SUA nor the predicted interaction. We believe a crucial flaw of this study was that telling participants how much would be needed to save one child beforehand undermined the point of SUA or CUA because participants could already multiply without using any specific asking technique. This is also why we did not include it in the main part of the manuscript.

B.2 Pilot 3 - Priming Scope Sensitivity

B.2.1 Methods

The aims of this study were twofold. First, to do a more fine-grained analysis of the Sequential Unit Asking technique. We varied the number of steps and investigated how that affects the amount that participants donate after the final step. In addition, we varied the difference between steps to see how this impacts the WTD judgements. Second, we tested whether, after being consistent for several small steps, people are primed to be consistent, thus making them more consistent in a later larger step.

1. Consistency priming: 1, 2, 3, 4, 5, 100 [6 steps]
2. Equal spacing: 1, 20, 40, 60, 80, 100 [6 steps]
3. Max steps: 1, 2, 3, 4, 5, 20, 40, 60, 80, 100 [10 steps]

Our first hypothesis was the people would give the most in the Max steps condition because there is some increase with each step. Our second hypothesis was that people give more in the consistency priming condition than in the equal spacing condition due to the priming idea discussed above. Our third hypothesis was that people would give more in the next-to-last step in the equal spacing condition (where there are 80 beneficiaries) than in the consistency priming condition (where there are five beneficiaries). This would show that conditional on the number of steps being fixed (and the size of all steps being the same), the participants will donate more the greater the size of each step is; in other words, they are at least somewhat scope sensitive.

Analyses

We used a Bayesian ANOVA in the package BRMS as in the previous study. Because we cannot meaningfully analyze scope insensitivity for participants that donated 0 (as they would donate 0 for any scope under scope consistency as well as scope

inconsistency), we excluded these participants. As preregistered, we excluded unbelievably large WTD judgements of more than \$10,000. As our data was (after excluding the zeros) expected to be positive and quite skewed, we used a lognormal likelihood rather than the more common Gaussian likelihood. We used a prior of normal(3,1) on the intercept, normal(0, 0.4) on the scope (as we expected scope insensitivity) and normal(0.4, 0.4) on the main effect of Asking Type, normal(1, 0.5) on σ . The reason why the priors are increasingly more narrow is that given the log-normal likelihood, the prediction on the linear scale corresponds to the exponent of the marginal mean. Therefore, using similar priors on intercepts and slopes would result in the prediction of an extremely large difference between the groups. We tested for the presence of the hypothesized mean differences by comparing models that include the interaction to models that do not include the interaction using Bayes factors. In addition, we used non-parametric Wilcoxon tests to check the robustness of our conclusions.

Participants

We recruited American participants through positiy. The study was run on 18th of February 2021. Participants received a payment of \$0.36 to participate in a 3 minute study. Initially, 201 participants signed up, we excluded 2 for indicating implausible age. In addition, we excluded 1 participant who donated more than \$10,000. We also excluded two participants who indicated impossible ages. This left us with a final sample of 177 participants. Mean age was 43.39 (sd = 12.06). 91 participants were female and 108 were male. This study received ethical approval by Harvard University's ethics review board.

B.2.2 Results and Discussion

As all conditions in this study were SUA conditions, we directly proceeded with excluding WTD judgements of 0 (21 participants) and more than 10,000 (1 participant). We only found weak evidence that more steps lead to higher WTD judgements ($bf_{10} = 2.12$). Hypothesis two was also not supported with only weak evi-

dence that consistency priming improved performance over equal spacing ($bf_{10} = 1.97$). Finally, we also found strong evidence that participants donated more for 80 in the equal spacing condition than for 5 in the consistency priming condition ($bf_{10} = 23.83$). This is also corroborated by non-parametric Wilcoxon test (Hypothesis one: $W = 1585.5$, $p = .1276$; Hypothesis two: $W = 2800.5$, $p = .095$; Hypothesis three: $W = 1307.5$, $p = .005$).

B.3 Prior Predictives for BRMs model

When sampling only from priors, we obtain predictions for the marginal means as in Figure B.1. In other words, we would estimate a donation of 37 for the DA condition and a donation of 56 for the unit asking condition on a low scope. Because of the hypothesized interaction of Asking Type and scope under H_1 , we predicted a donation of 71 for unit asking and scope 10,000. The skew of the credible intervals is a feature of the lognormal likelihood, which helps accommodate strong positive outliers.

B.4 Share of Participants Donating for DA, CUA, and SUA conditions in The Different Studies

Table B.1: Median WTD judgements For The Seven Conditions of Study 2

Condition	Scope	Median	% Donating
DA	100	\$15	36.59
CUA	100	\$40.0	17.28
SUA	100	\$79.0	14.81
DA	10,000	\$30.0	30.00
CUA	10,000	\$50.0	23.75
SUA	10,000	\$75.0	15.19
SUAI*	10,000	\$50.0	27.85

Note. *SUA scaling up with increase per step constant.

B.5 Step-by-Step Guide to the Bayesian Analysis

This Appendix contains a step-by-step guide to the Bayesian Analysis executed in this paper, including R code. All analyses use R version 4.1.2, and the packages

Figure B.1: Prior Predictives for the Lognormal Model.

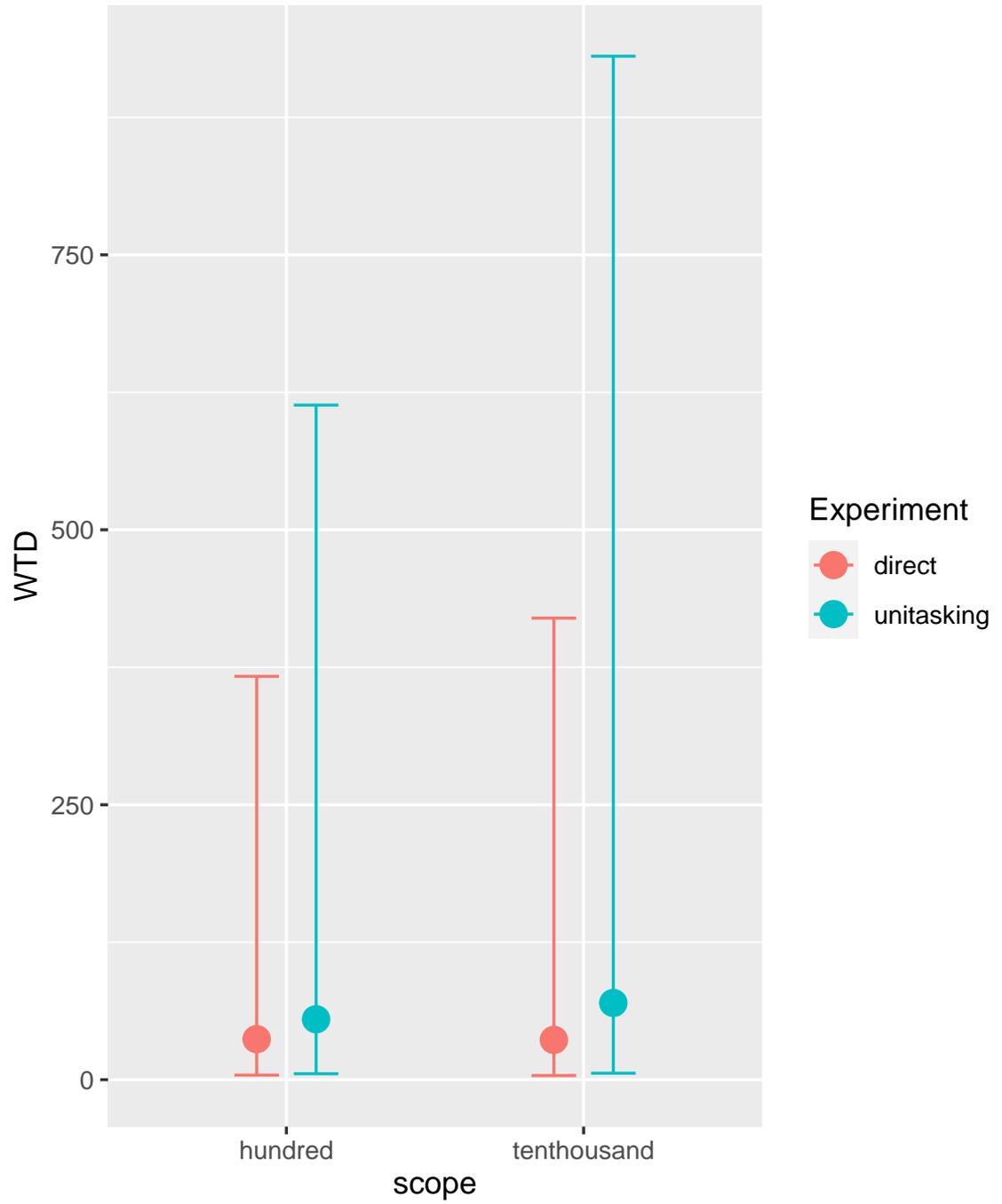


Table B.2: Median WTD judgements For The Seven Conditions of Study 3

Condition	Scope	Median	% Donating
CUA	10	\$50.0	9.76
SUA	10	\$50.0	15.48
SUA	10 & scope known	\$50.0	11.39
CUA	50	\$40.0	8.43
SUA	50	\$50.0	14.63
SUA	50 & scope known	\$50.0	10.84
SUAI*	50	\$75.0	14.81

Note. *SUA scaling up with increase per step constant.

brms version 2.16.3, rstan version 2.21.3, and BayesFactor version 0.9.12. This Appendix is merely intended as a conceptual explanation for the most important parts of the R code; for a more general accessible introductions to Bayesian statistics, see Wagenmakers (2020) and Wagenmakers et al. (2016).

B.5.1 Step 1: Testing Whether Proportion of Participants Donating in the First Place is Affected

After reading in the data (and excluding participants that failed attention checks), we use the `contingencyTableBF` function from the `BayesFactor` package to check whether the intervention influenced the share of participants donating in the first place.

```
#Construct Matrix of Counts
counts <- c(sum(na.omit(subset(data, Experiment..Pilot. == "Control")$WTD)
> 0),
sum(na.omit(subset(data, Experiment..Pilot. == "Control")$WTD) ==
0),
sum(na.omit(subset(data, Experiment..Pilot. == "Intervention1")$WTD)
> 0),
sum(na.omit(subset(data, Experiment..Pilot. == "Intervention1")$WTD)
== 0),
sum(na.omit(subset(data, Experiment..Pilot. == "Intervention2")$WTD)
> 0),
sum(na.omit(subset(data, Experiment..Pilot. == "Intervention2")$WTD)
```

```

== 0))
counts <- matrix(counts, nrow = 2)
#inspect proportion of people donating 0
counts[2,]/colSums(counts)
#Check whether it depends on the manipulation
BayesFactor::contingencyTableBF(counts, sampleType = "jointMulti")

```

B.5.2 Step 2: Fitting the Models

Once we established that the proportion of participants donating in the first place is not affected, we can go ahead and fit the models. In order to test for the presence of a difference between groups, we specify one model that assumes a difference between groups (\mathcal{H}_1) to one model that does not assume a difference (\mathcal{H}_0).

```

#H1
data.brms1 = brm(bf(WTD ~ Experiment..Pilot., center = FALSE),
data = data, iter = 10000, warmup = 2000, chains = 4,
cores = 12, thin = 2, refresh = 0, family = lognormal(),
prior = c(prior(normal(0.4, 0.4), class = "b"),
prior(normal(3, 1), class = "b", coef = "Intercept"),
prior(normal(1,0.5), class = "sigma")
), save_all_pars = T)

#H0
data.brms0 = brm(bf(WTD ~ 1, center = FALSE),
data = data, iter = 10000, warmup = 2000, chains = 4,
cores = 12, thin = 2, refresh = 0, family = lognormal(),
prior = c(prior(normal(3, 1), class = "b", coef = "Intercept"),
prior(normal(1, 0.5), class = "sigma")
), save_all_pars = T)

```

We can see that this code specifies two competing models, the \mathcal{H}_1 includes an effect of the intervention ($WTD \sim \text{Experiment..Pilot.}$), whereas the \mathcal{H}_0 model does not ($WTD \sim \text{Experiment..Pilot.}$). Further, we specify the number of

iterations and warmup for the MCMC sampling (details on MCMC sampling and checking of model convergence are beyond the scope of this guide, see, for example, McElreath, 2020) and prior distributions. The priors here were based on pilot data. There is ample guidance available on different approaches to specifying prior distributions, and we refer the reader to the literature (e.g., Stefan et al., 2022). Finally, we specified `family = lognormal()`. This indicates that we are relying on a lognormal likelihood rather than a normal likelihood.

Step 3: Comparing the Models

Finally, we compare the two models under consideration using Bayes Factors:

```
bayes_factor(data.brms1, data.brms0)
```

This code will use bridge sampling to estimate the Bayes factors (more details in Gronau, Sarafoglou, et al., 2017). The Bayes factor compares the likelihood of the data under the two competing models. Here we calculate the Bayes factor in favor of the alternative (as `data.brms1` is listed first in the brackets). As a rule of thumb, Bayes factors between 1 and 3 are considered weak evidence, Bayes factors between 3 and 10 are considered moderate evidence, and Bayes factors larger 10 are considered strong evidence (M. D. Lee & Wagenmakers, 2013). When the evidence for the null is considered, the Bayes factor is simply inverted ($BF_{10} = 1/BF_{01}$).

Appendix C

Decisions Under Extinction Risk

C.1 Existing Paradigms and Their Limitations

Among the most popular paradigms (for a recent overview, see Pedroni et al., 2017), the Balloon Analogue Risk Task (BART) stands out as including something resembling an extinction event (Lejuez et al., 2002, 2003; Pleskac, 2008; Pleskac et al., 2008; Wallsten et al., 2005). In the BART, participants' earnings increase each time they pump a balloon. Each additional pump, however, risks bursting the balloon and earning nothing. While the bursting of the balloon might be viewed as an extinction event, the BART does not capture how people reason about the types of extinction risks described above. This is because the losses are limited to that specific balloon trial, leaving the earnings from past – and future – balloons unaffected. Also, the BART imposes a strong relationship between choice and event probabilities: the more you pump, the greater the chances that the balloon will pop at each subsequent pump. However, in many real-life examples, the frequency of committing a risky act does not necessarily increase the risk of extinction with each subsequent action. That is, *ceteris paribus*, your 20th jaywalk is not more risky than your 1st.

The Bomb Risk Elicitation Task (BRET, Crosetto and Filippin, 2013) is another paradigm worth highlighting. In this task, participants decide how many boxes to collect out of 100. One of these boxes contains a bomb. Participants receive more earnings with each box they collect, but if they happen to collect the box containing the bomb, their earnings are reduced to zero. The BRET can be administered in a

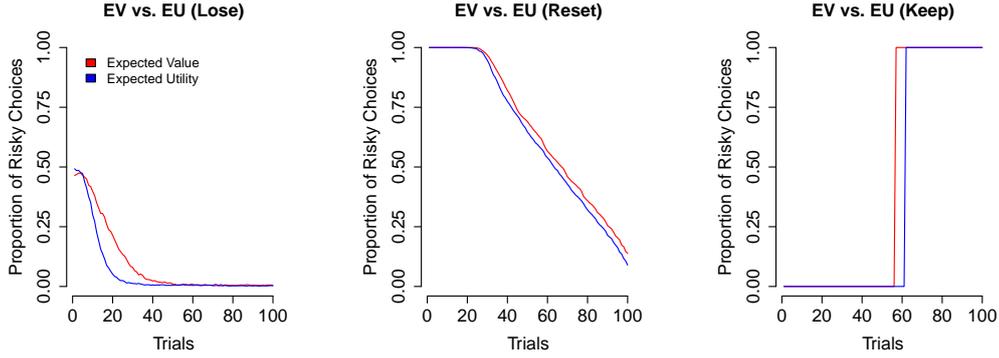
static and a dynamic version. In the static version, people specify the number of boxes to collect in advance. In the dynamic version, one box is collected every second until the participant presses the stop button. While the BRET has been shown to be a useful tool to study risky choices in general, it was not designed to apply to extinction risks and differs from decisions involving extinction risks in at least two respects. First, exactly one extinction outcome is among the set of options with certainty. This, again, results in a dependency between the number of boxes selected and the probability of finding the bomb, similar to the BART. Second, in the BRET, participants either choose the number of boxes they wish to collect in advance, or observe boxes being collected sequentially (without making an active decision on each instance) until they decide to hit the stop button. Both of these versions are quite different from real-life decisions about extinction risks, such as jaywalking, which involve repeated choices between different options..

Because of the growing interest in human action in the face of extreme risks (for theoretical papers see e.g., Slovic & Weber, 2013; Sundh, 2024; Weber, 2006), it is unsurprising that the need for experimental research has been recognised: Pfors and Van Dam (2018) proposed a new choice paradigm in which people make repeated choices under the risk of extreme *black swan* events that completely wipe out their accumulated earnings. Specifically, participants first executed a single choice and then (without feedback from their first choice) specified a policy according to which they would choose 2000 times in a row. The riskier lottery included a small-probability event that wiped out all previous earnings. However, these black swan events cannot be considered to be extinction events as defined in this paper (where losses are unrecoverable), given that participants were still able to accumulate future gains. Also, the initial commitment to a specific number of risky choices prevents any influence of experience and the investigation of how people frame and adjust their choices across trials (e.g., Bland, 2019; Rabin & Weizsäcker, 2009; Read et al., 2000). Further, Elga et al. (2024) discuss resource allocation in the face of extinction risks; however, their results mostly apply to scenarios where the risk is very high, and the number of repeated choices is low, which is different from the

setup proposed in our paper and extinction risks in the real-world.

C.2 Expected Value vs. Expected Utility Optimal Strategies for the Three Conditions

Figure C.1: Optimal Strategies According to Expected Value vs. Expected Utility



Note. Payoffs, probabilities and trial number are the same as in Experiments 1 and 2. We used an exponential function to model the diminishing returns with an exponent of 0.66 based on Kellen et al. (2016).

C.3 A Priori (Non-Dynamic) Optimal Strategy for the Lose Condition

What if people do not take the current endowment into account to dynamically adjust the optimal strategy but instead commit to a single number of risky choices in advance, which they play? We can model this type of *a priori* expected value of different numbers of risky choices in the complete extinction case using Equation C.1,

$$\begin{aligned}
 EV_{\text{a priori}}(N_{\text{risky}}) = & \underbrace{(\bar{r}_{t, \text{safe}} \times (N_{\text{total}} - N_{\text{risky}}))}_{\text{Expected value of safe trials if survival}} + \underbrace{\bar{r}_{t, \text{risky}} \times N_{\text{risky}}}_{\text{Expected value of risky trials if survival}} \\
 & + \underbrace{Z}_{\text{Endowment}} \times \underbrace{p(s)^{N_{\text{risky}}}}_{\text{Total probability of survival}}.
 \end{aligned} \tag{C.1}$$

$\bar{r}_{t, \text{safe}}$ denotes the expected value of choosing the safe lottery. $\bar{r}_{t, \text{risky}}$ denotes the expected value of choosing the risky lottery (assuming the player does not go

extinct). N_{total} denotes the total number of trials of the experiment. N_{risky} denotes the number of risky trials being played. $p(s)$ denotes the probability of surviving (i.e., $1 - p(\text{extinction})$) when playing the risky gamble. Finally, Z denotes the endowment. When starting in the first trial with zero endowment this term is simply set to zero.

Based on this expected value the optimal a priori number of risky choices is simply

$$\operatorname{argmax}_{N_{\text{risky}}} \text{EV}_{\text{a priori}}(N_{\text{risky}}). \quad (\text{C.2})$$

Notably, the difference in terms of expected payoff between the dynamic and the a priori strategy is relatively small. For the payoffs and probabilities used in Experiment 1, the expected payoff following the dynamic strategy would be 57.87p, whereas the expected payoff following the a priori strategy would be 57.05p.

C.4 Estimating False Positive Rate and Power

To estimate the statistical power of our design and to verify that our analysis would indeed accurately capture the data generating fixed effects between conditions in the presence of selection effects, we conducted a simulation-based power analysis for mixed effects models. While the R script shared on the OSF page can easily be adapted for a variety of settings, this appendix focuses on the setting of an interaction between trial number and extinction condition as well as a main effect of extinction condition (similar to Experiment 1). We simulated from a trial number slope, which starts from a probability of selecting the risky option of .5 in the first trial, declining to .05 during the experiment in Condition A, while it increases from 0.05 to 0.5 in Condition B. Further, we added a main effect of condition, whereby condition B is shifted by 1 on the logit scale (in comparison, the condition effect found in Experiment 1 was 1.12). We simulated from models without this main effect to assess the false positive rate and from models with this main effect to assess the statistical power. The random intercepts and slopes of trial number were 1 on the logit scale. We assume a sample size of 80 per condition. We also compared the mixed effects model approach to simply t -testing the difference in the proportion of

risky choices between conditions. We repeated each condition 1000 times.

Table C.1 summarises the results. The first two columns show the performance of both models when there is no selection. This indicates an almost nominal error rate, calibrated estimates and good power to detect an effect. The second two columns show what happens if we introduce selection effects (i.e., a chance of dropping out when choosing the risky option). The performance of the mixed effects models is similar, with slightly lower power due to the reduced number of trials. We can also see a slightly above nominal error rate, likely due to non-convergence of some models. In the manuscript, we always compare to a reduced, converged random effects structure in this case. However, merely *t*-testing the difference in proportions between conditions dramatically increases the error rate as the selective dropout distorts the simple proportion estimates (unlike the mixed model, which can account for this through modelling the trial number effect) and actually has a higher chance of claiming an effect when there is none than when one is present (due to the different starting points from the fixed slope and the condition effects cancelling each other out under selection).

Table C.1: Power Analysis for Mixed Model and T-Testing Proportions

Selection	No		Yes	
	0	1	0	1
True condition difference (logit)	0	1	0	1
Power mixed model	3.9%	¿99.9%	6.4%	¿99.9%
Power difference between proportions	4.1%	¿99.9%	¿99.9%	73%
Estimate mixed model (probability)	0.00*	0.16*	0.01*	0.18*
Estimate difference between proportions	0.00*	0.16*	0.22*	-.10*

Note. Estimates indicated with * are on linear probability scale rather than logodds scale.

C.5 Full Model Specification and Implementation

C.5.1 Full Model Specification

Our model had the three following response models that describe the probability of choosing risky ($p(r)$) either in terms of a constant probability or as a function of the current trial number (t):

$$P(\text{r—safe st.}) = \text{logit}(\alpha_{s,i})^{-1} \times 0.2 \quad (\text{Safe State}) \quad (\text{C.3})$$

$$P(\text{r—reg. st.}) = \text{logit}(\alpha_{reg,i} + \beta_{reg} \times N_{\text{trial}})^{-1} \quad (\text{Logistic Regression State}) \quad (\text{C.4})$$

$$P(\text{r—risky st.}) = \text{logit}(\alpha_{r,i})^{-1} \times 0.2 + 0.8 \quad (\text{Risky State}) \quad (\text{C.5})$$

N_{trial} denotes the trial number. We used a hierarchical implementation for α_s , α_{reg} , α_r , and β_{reg} with random intercepts and random slopes. We did not directly model the covariation between intercepts and slopes, as this would increase the number of parameters that need to be estimated, considerably, impeding model convergence.

We then introduced a hidden Markov model component that allowed transitions between the safe and the risky states to model the switch points. However, as described above, we did not allow switches to and away from the regression state. The transition matrix below shows the structure of the model with the rows and columns being ordered as (1) safe, (2) regression, (3) risky and ω denoting the probability of staying in the safe or risky state once in it:

$$\Gamma_{i,j} = \begin{pmatrix} \omega & 0 & (1-\omega) \\ 0 & 1 & 0 \\ (1-\omega) & 0 & \omega \end{pmatrix} \quad (\text{C.6})$$

Finally, the starting probability for each state is denoted by the uniform three-dimensional simplex ρ with a separate ρ estimated for each participant (i.e., no hierarchical structure for ρ).

Prior Distributions

For the grand mean and standard deviation on all α and β we always use a normal and half-normal distribution:

$$\mu \sim \text{normal}(0,1) \quad (\text{C.7})$$

$$\sigma \sim \text{normal}(0,1)_+ \quad (\text{C.8})$$

For the transition probabilities simplex $[\omega, 1 - \omega]$ we use

$$[\omega, 1 - \omega] \sim \text{Dirichlet}(30,1), \quad (\text{C.9})$$

which is equivalent to

$$\omega \sim \text{beta}(30,1), \quad (\text{C.10})$$

and reflects the assumption that participants who are in a certain state (i.e., play mostly risky or mostly safe) are more likely to play risky or safe again on the next trial than to change the state. Finally, the initial state probability has the uniform prior

$$\rho \sim \text{Dirichlet}(1, 1, 1) \quad (\text{C.11})$$

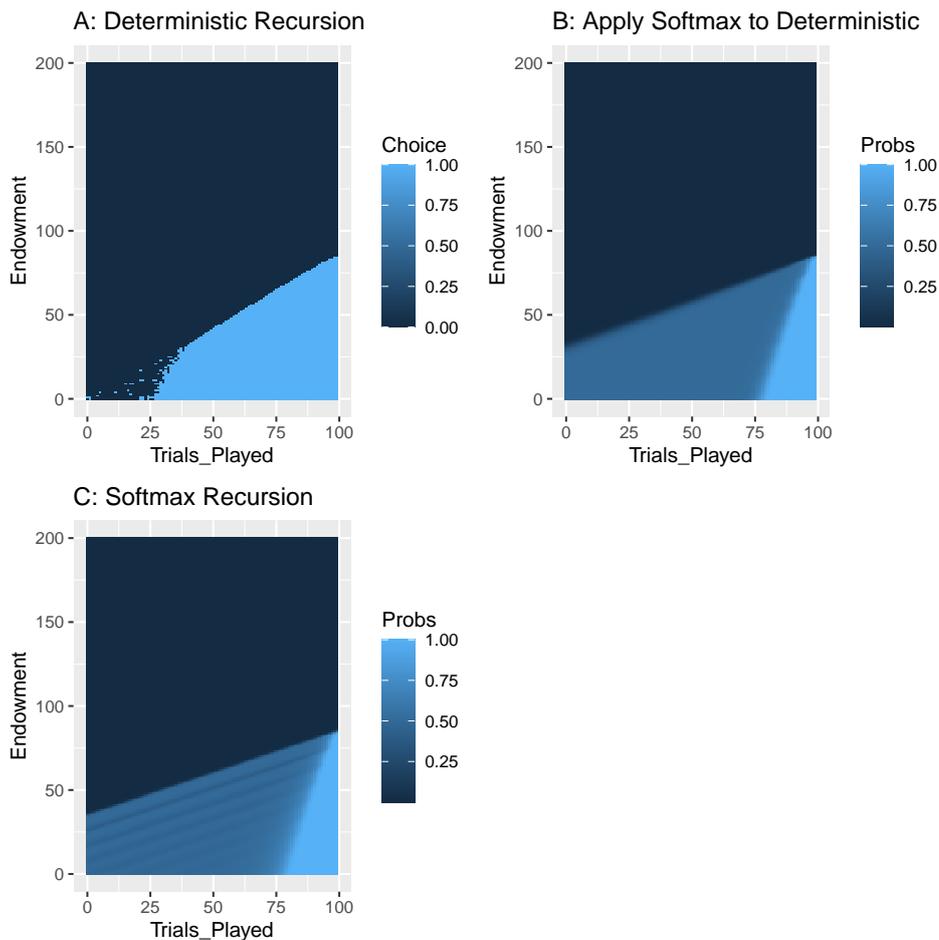
Implementation

We implemented the model in `cmdstan` (version, 2.34.1 cite) and fitted it using `cmdstanr` (Version 0.6.1, cite). We ran each model with 1000-3000 warmup iterations and 2000-6000 sampling iterations (for Experiments 1 and 3 we ran 3000 iterations, of which 1000 were warmup iterations using `adapt delta = 0.95`. For Experiment 2 we run 9000 iterations with 3000 warmup iterations using `adapt delta = 0.98`). For Experiment 4, we run 9000 iterations, 3000 warmup with `adapt delta .95` in the Lose condition and `adapt delta .98` in the Keep condition). All \hat{R} were smaller

1.03.¹

C.6 Heatmaps of Optimal Choice as a Function of Endowment and Trial Number

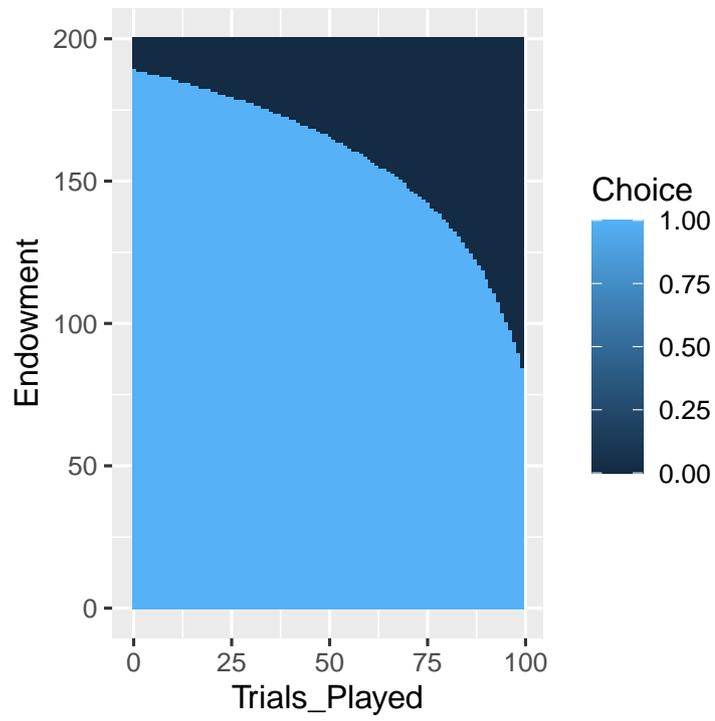
Figure C.2: Heatmaps Visualizing the Optimal Choice As a Function of Trial Number and Money for Experiment 1 - Lose Condition.



Note. Panel A shows a striking and surprising pattern, where there is no clear boundary between playing safe and risky choices. Panels B and C help us understand the reason for this by using a softmax transformation to derive choice probabilities based on the expected value of the two options with a very low temperature of 0.02. Panel B applies a softmax transformation to the deterministic recursive function, while Panel C uses a recursive softmax function directly. Comparing both of these panels to Panel A suggests that in a large area of the parameter space, there is almost no difference between playing safe and playing risky in terms of the dynamic solution. So the pattern of Panel A is the result of only tiny differences in EV between the safe and risky option that can flip based on how sequences of choices starting from a certain tile make it easier to reach the safe vs. risky state.

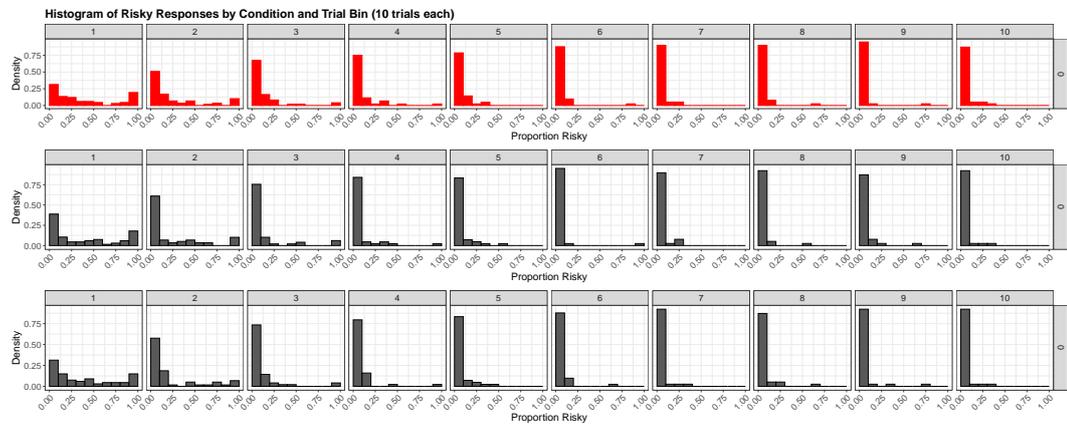
¹For Experiment 1 all Rhats were smaller 1.01, for Experiment 2 all Rhats were smaller 1.03, for Experiment 3a all Rhats were smaller 1.01, for Experiment 3b all Rhats were smaller than 1.02.

Figure C.3: Heatmap Visualizing the Optimal Choice As a Function of Trial Number and Money for Experiment 2 - Reset Condition.



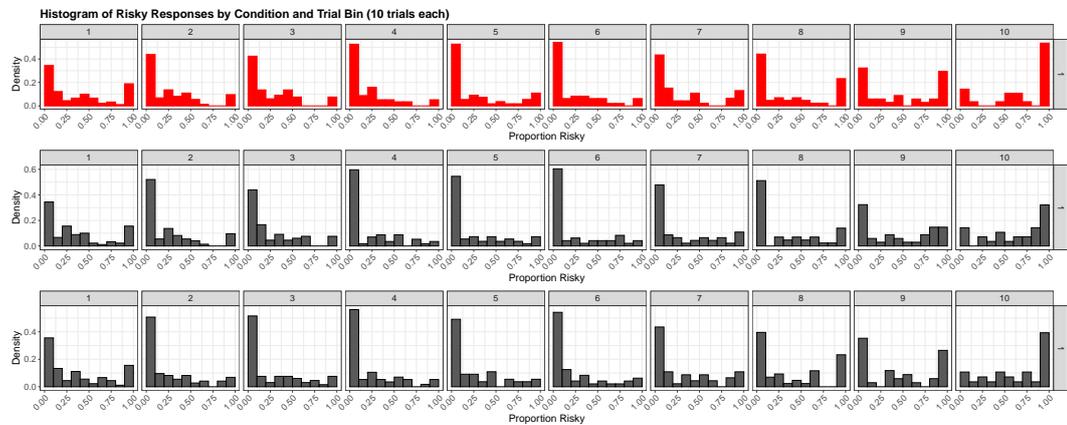
C.7 Comparison of Posterior Predictions and Data

Figure C.4: Model Predictions vs. Data in the Lose Condition (Exp. 1)



Note. Histograms show the distribution of risky choices within ten trials. For example, panel “1” shows the distribution of average probabilities to make a risky choice across participants for trials 1 to 10. The top row indicates the data and the rows 2 and 3 below each show one random draw from the posterior predictive distribution.

Figure C.5: Model Predictions vs. Data in the Keep Condition (Exp. 1)

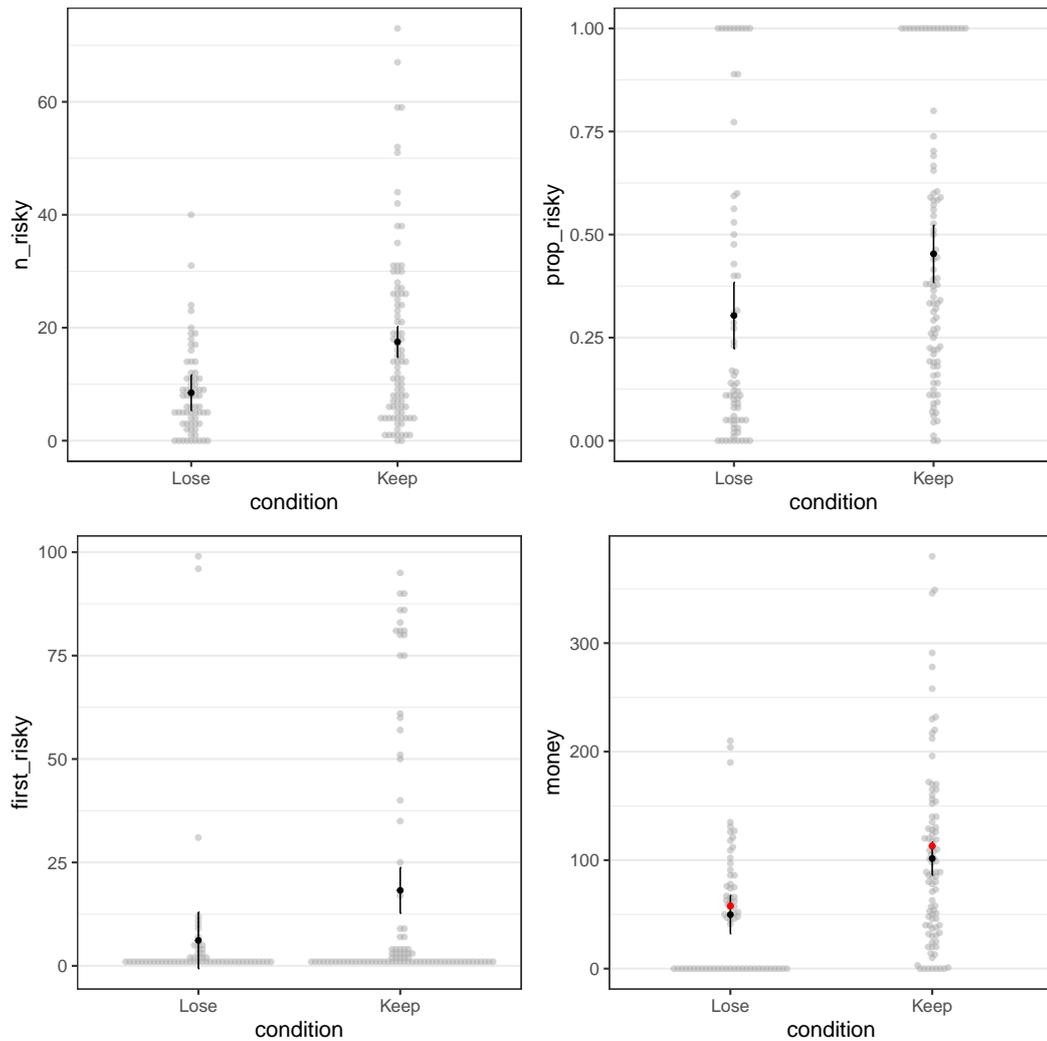


Note. See Figure C.4.

C.8 Additional Results

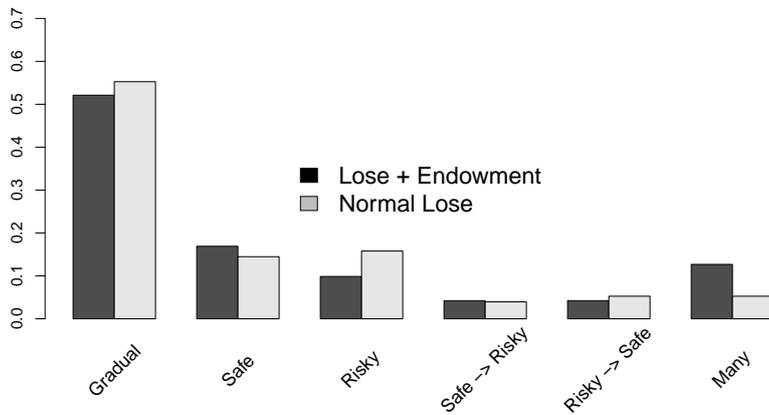
C.8.1 Visualization of Additional Variables for Experiment 1

Figure C.6: Number of Risky Choices, Proportion of Risky Choices, Trial Number of First Risky Choice, and Money Earned Between Conditions for Experiment 1



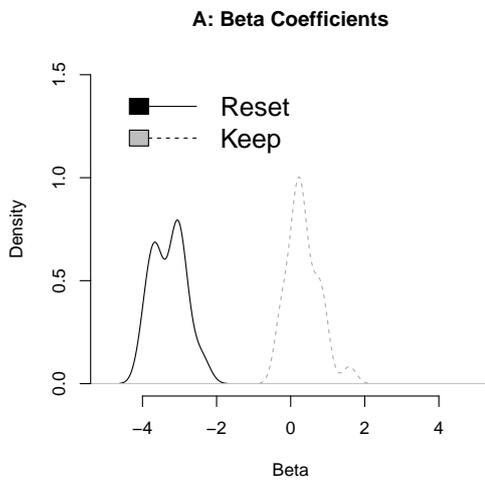
Note. The red dots in the bottom left panel indicate how much money an agent following the optimal strategy would make on average.

Figure C.8: Introducing Endowment and Losses Does Not Affect the Strategies Participants Employ



C.8.2 Beta Coefficients for Experiment 2

Figure C.7: Beta Coefficients of Those Participants Assigned to the Gradual Strategy.



C.8.3 Computational Modelling Results for Experiment 3.

C.8.4 Computational Modelling Results for Experiment 4.

Figure C.9: Proportion of Strategies for the Two Different Maximum Scopes in the Keep condition

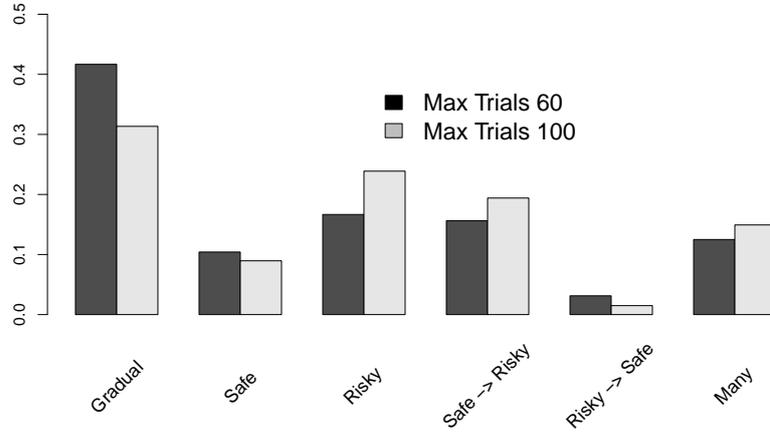


Figure C.10: Proportion of Strategies for the Two Different Maximum Scopes in the Lose condition

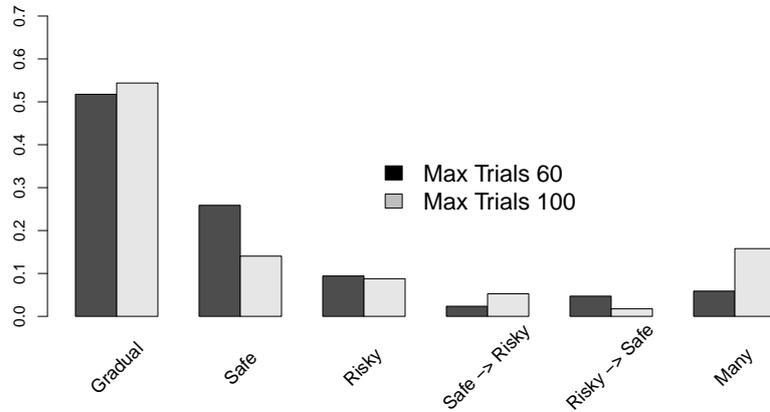
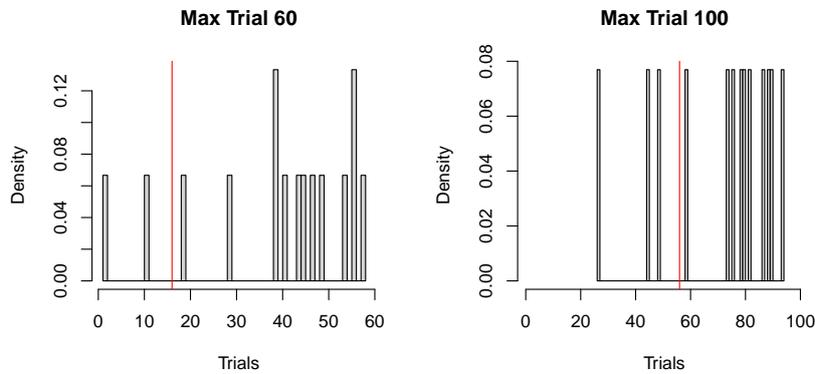


Figure C.11: Switch Points in Comparison to the Optimal Switch Point in the Keep Condition



C.9 Optimal Strategies Collective Game

C.9.1 Collectively Optimal Solution Game

A Priori Strategy

The a priori strategy would be the same for an individual game and a group game with the same number of total trials in terms of the collective expected value. In other words, a group with 5 people that play 20 trials each could make the same amount of money as one individual playing 100 trials.

Dynamic Strategy

In the group game, participants play five choices in each round rather than one and the total number of rounds is reduced to 20 (giving 100 total choices as in the individual game). This implies that they have slightly less ability to take their luck in previous decisions into account in comparison to the individual game. For example, if someone played 4 trials risky in the individual game, they can then take their earnings on the first four trials into account before making the next decision. In contrast, in the collective game the first five group choices are played simultaneously and participants trying to make an optimal decision for the group can only update based on their luck in previous trials every 5 choices.

Table C.2: Possible Combinations of Survival Outcomes and Their Associated Probabilities for Five Players, which Play four Risk Choices

$N(0 r)$	$N(50 r)$	$N(0 s)$	$N(5 s)$	$p(outcome)$
0	4	0	1	0.02545332
1	3	0	1	0.10181328
2	2	0	1	0.15271992
3	1	0	1	0.10181328
4	0	0	1	0.02545332
0	4	1	0	0.02545332
1	3	1	0	0.10181328
2	2	1	0	0.15271992
3	1	1	0	0.10181328
4	0	1	0	0.02545332

Note. $N(0|r)$ =number of +0 outcomes from the risky lotteries, $N(10|r)$ =number of +10 outcomes from the risky lotteries & $N(0|s)$ =number of +0 outcomes from the safe lotteries & $N(1|s)$ = number of + 1 outcomes from the safe lotteries & $p(outcome)$ = probability associated with that outcome. The probabilities here do not sum to one as the table only shows the survival outcomes and not the outcomes where people go extinct as in this case the solution would just return 0 (the p column sums to 0.814 and the probability of extinction would be $1 - (0.95)^5 = 0.186$ so taking the probability of extinction into account the probabilities indeed sum to one).

In the group game, we therefore evaluate a recursive function for 20 trials, whereby in each trial participants can chose between playing 0 to 5 risky choices.

First, we create a matrix which denotes all possible survival outcomes and their associated probabilities for different numbers of risky choices. This can be done easily in R using the `expand.grid()` function. For example, in the five player game and for 4 out of 5 risky choices this would be given in Table C.2

We can then calculate the expected value for each number of risky choices by summing the rewards over each possible outcome weighted by their probabilities. For example the expected value of playing four risky choices would be:

$$EV_{col,4}(N, \epsilon) = \sum_j p(r_j) \times EV_{col}(N - 1, \epsilon + r_j), \quad (C.12)$$

where j is an index that loops over all rows in Table C.2. When zero trials are left

the function again returns the endowment

$$EV(0, \varepsilon) = \varepsilon, \tag{C.13}$$

and the optimal decision is the one with the largest expected value

$$EV(N, \varepsilon) = \max(EV_{\text{col}, 0:5}(N, \varepsilon)) \tag{C.14}$$

Taking Into Account Coordination Problems for the Group Dynamic Strategy

The previous optimal strategy still assumes that the group can play whichever number of risky choices is optimal based on the dynamic solution. This should in principle be feasible in the median voting condition; however, in the other conditions precisely playing the collectively optimal number might be difficult given that players are not able to communicate. For instance, if it is optimal to play 3 out of 5 risky choices in the next round, it is likely that players are not able to coordinate on three of them playing risky and instead play, for example, 5 out of 5 risky choices or just 2 out of 5. In other words, even if players if they were trying to maximise the collective good and had full understanding of the optimal strategy, they may not be able to coordinate well enough to actually execute it as a group. These coordination problems would only arise when the optimal number of risky choices is between 1 and 4, when all 5 choices should be risky or all 5 choices should be safe it would in principle be possible to coordinate on that as there is no difficulty of allocating who should play the risky and who should play the safe choices. To estimate what payoffs could realistically be achieved given the presence of coordination problems, we modify the script for simulating the optimal strategy so that whenever the optimal number of risky choices is not 0 or 5, the group plays a randomly drawn number of risky choices from a uniform distribution between 0 to 5 rather than the optimal number. In practice, participants may be more responsive to the optimal number of risky choices, even when it is between 0 and 5 (i.e., when 1

risky choice is optimal, probably fewer participants would play risky than when 4 are), and this implementation may therefore be considered a lower bound on how much very strong coordination problems could distort the results.

Comparison Of Optimal Proportion of Risky Choices and Payoffs Between Strategies

How do the payoffs and the proportion of risky choices between the different optimal strategies outlined above compare for the game studied in this paper? Table C.3 compares the a priori solution, the dynamic individual solution, the dynamic group solution, and the dynamic group solution taking possible coordination problems into account. The main takeaway from this table is that the proportion of risky choices and the payoffs between these different strategies do not differ substantially. Further, we can see that the a priori solution is a good approximation to the dynamic solutions, with the payoff being less than 1p lower for this approach in comparison to the dynamic solution.

Table C.3: Comparison Between Different Optimal Strategies in Terms of Payoffs and Proportion Risky

Strategy	Average Riskiness	Payoff
A priori	8%	57.05p
Dynamic Individual	8.93%	57.88p
Dynamic Group	8.87%	57.87p
Dynamic Group + Coordination Problems	10.77%	57.56p

Note. All simulations are based on the payoffs and probabilities used in the present experiment (see Method section). The proportion of risky choices for the dynamic strategy is based on simulating XXX agents playing the game.

C.9.2 Individual Solution in the Collective Game

The individually optimal strategy in the collective game is essentially the same as the solution to an individual game with only 20 rounds. In such a game, the relative costs of extinction are much lower (since there is less that can be gained in 20 trials than 100 trials). Therefore, the optimal proportion of risky choices is much higher than in an individual game with 100 trials. By contrast, the background level of

Table C.4: Individually Optimal Strategy in the Collective Game

Number of other players being risky	Individual Average Riskiness	Expected Payoff
0	84%	182.23p
1	84%	63.62p
2	85%	23.99p
3	84%	8.99p
4	82%	2.70p

risk (i.e., how risky other people in your group are), does not play a role for the proportion of risky choices in the optimal individual strategy as higher background risk both reduces the cost of extinction and the benefits of non-extinction equally. However, the background risk of course has considerable impact on the expected payoff, such that the expected average individual earnings are lower the higher the background risk.

Table C.4 compares the individually optimal strategy (i.e., dynamic individual solution) for the game in this paper with four different levels of background risk, which are based on 0 to 4 of the other players in the group playing risky each round. This indicates that if everyone else plays safe, a self-interested player, who is more risky can get a substantially higher payoff by playing more risky than the payoff of following the collectively optimal strategy (84% risky with an expected payoff of 182.23p versus 9% risky with an individual payoff of 57.87%). However, as other players also start to deflect the expected payoff falls below the expected payoff that could be achieved by following the collectively optimal strategy to as low as 2.7p. Note that the slight differences in the proportion of risky choices are due to numerical variability in the simulation (each game setting was simulated 5000 times).

C.10 Details on Estimated Average Riskiness, False Positive Rate, and Power for the Collective Task

C.10.1 Estimated Average Riskiness

If a group goes extinct during the experiment, their data from after the extinction event is not used for the analysis. This may distort the inferences as some groups drop out earlier than others and dropping out is related to the riskiness of their choices. We can account for this by comparing marginal means estimated from the mixed effects model giving equal weight to each trial number (rather than simply calculating the mean proportions within each group). This procedure accounts for selective attrition as it is based on model-predictions and thus for those participants that dropped out earlier the counterfactual slope of how they would have chosen in later trials (based on their previous choices and the hierarchical shrinkage from the other participants) is taken into account. In the next section, we include a simulation study, which demonstrates that our analysis is indeed robust to selection effects, and shows how other procedures, such as merely *t*-testing the proportion of risky choices can be severely distorted in the presence of selection effects. Finally, because the adjustment is based on model estimates, it is appropriate only to the extent that the model captures the data well. Therefore, we also present a comparison of model predictions versus data within each group in the subsection ‘Model Predictions vs. Data’.

C.10.2 Demonstrating the Influence of Selection Effects and the Robustness of Estimated Average Riskiness

To maximise the potential influence of selection effects on the design, let us consider the following setup for the data-generating process: In condition A participants start by playing risky 30% of the time and reduce their probability of playing risky to 5% in the last round. By contrast, in condition B participants start by playing risky 5% of the time but increase their riskiness to 30% in the last round. Without any selection the average riskiness between these two conditions would be the same. How-

ever, when selection effects are introduced participants in condition A will appear more risky than condition B. The reason is that if participants in group A drop out during the task their estimated riskiness is higher than what it would have been had they played all rounds because they do not get to play the last trials in which they would have been mostly safe. For condition B the same type of reasoning implies underestimation when selection effects are introduced.

Testing mean differences within the mixed effects model adjusts for this by explicitly modeling the slope of choices as participants proceed through the task. To test the robustness of our analytical approach we simulate 1000 tests of the condition difference for two groups of participants following the choice patterns of conditions A and B in the previous paragraph. To assess the influence of selection, we compare the estimates and tests in a scenario with selection effects to the estimates and tests in a scenario without selection effects. We assume 50 groups per condition and a standard deviation of 1 on the logit scale for random intercepts and random slopes. As this analysis focuses on the potential false positive inflating impact of selection effects, we do not assume any mean differences between the groups and focus on the false positive rate. For a power analysis based on a data-generating process more similar to what is likely in our design see the next section.

Table C.5 compares estimates and false positive rates for *t*-testing the mean difference in the proportion of risky choices versus using estimated average riskiness for estimation and testing within the mixed model. While the naive *t*-testing approach strongly inflates false positives (with a false positive rate of 83%) once selection effects are introduced, the mixed model retains an error rate very close to 5%. The influence on the estimates is in line with the influence on the tests: The estimated average riskiness is not meaningfully affected by selection effects, whereas the estimated difference between proportions changes considerably.

C.10.3 Power Analysis

For the power analysis, we use the same approach as in the previous section. However, we now assume parallel slopes from 0.3 to 0.05, which is more similar to the data we found in the pilot study. We also use 1.3 as standard deviation on the ran-

Table C.5: Influence of Selection Effects on False Positive Rate (FPR)

	No selection	Selection
FPR mixed model	4.1%%	4.7%%
FPR <i>t</i> -test	4.3%	87.3%
Mean Estimate mixed model	0.003	0.002
Mean Estimate difference between proportions	0.004	0.108

dom intercepts and 2.8 on the random slopes based on the pilot study. Further, we focus only on the setting in which selection effects due to extinction are present as that is the case in our study. We simulate from a condition difference of 0.7 on the logodds scale, which corresponds to a difference of $\approx 10\%$ on the linear probability scale. This indicates a statistical power of 95% with 70 groups per condition.

C.10.4 Model Predictions vs. Data

Figure C.12 visualizes the model predictions (red line) vs data for the pilot study. This shows that the model captures the main patterns in the data relatively well. Due to the low sample size, there is a large amount of round to round stochasticity in the data, which is not captured by the model. We would therefore expect an even better fit of the model to the data in the main study with more participants.

Figure C.12: Model Predictions (Red) vs Data for the Pilot Study

