The cognitive contours of punishment

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Psychology

have examined a dissertation entitled

The cognitive contours of punishment

presented by Arunima Sarin

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature [redacted]

Typed name: Prof. Fiery Cushman

Signature [redacted]

Typed name: Prof. Joshua Greene

Signature [redacted]

Typed name: Prof. Tomer Ullman

Signature [redacted]

Typed name: Prof. Rebecca Saxe

Signature [redacted]

Typed name: Prof.

Date: May 2, 2023

The cognitive contours of punishment

A Dissertation Presented

By

Arunima Sarin

To

The Department of Psychology

In partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

In the subject of

Psychology

Harvard University

Cambridge, Massachusetts

May 2, 2023

The cognitive contours of punishment

Abstract

People have many ways of influencing others, and one powerful method is through punishment. Punishment can be used to correct the misaligned values of those who commit moral infractions and to instill new values in those learning new things. We have a natural inclination to punish others and quickly learn from the punishments that we receive. In this dissertation I explore the cognitive mechanisms that allow us to learn from and teach with punishment successfully. Specifically, I ask how we conceptualize and structure punishments, how we learn from them, and how punishments modify our subsequent thoughts and behaviors. I examine two possibilities. The first suggests that we view punishments purely as a form of reinforcement and respond to it by using our cognitive ability for reinforcement learning. On this view, punishment is effective in changing behavior simply because it is inherently unpleasant. The second process suggests that we view punishment as a form of communication and engage with it by relying on our capacity for theory of mind. On this perspective, punishment serves as an informative signal conveying the punisher's values and goals. Its efficacy in modifying behavior stems from the message it conveys and the inferences it enables learners to draw. Across various chapters of this dissertation, I explore the distinct roles each of these processes plays in shaping our experiences with punishment. Chapter 1 shows that people generally engage with punishment through communicative inference, wherein teaching and learning from punishment entail expressing and inferring communicative intent. Chapters 2 and 3 identify scenarios where the direct and inherent unpleasantness of punishment plays a crucial part in eliciting the desired behavioral change.

Finally, Chapter 4 demonstrates that when imparting novel information to another person, instead of simply reinforcing good or bad actions, people structure their punishments communicatively to influence the learner's belief in a way that minimizes uncertainty and prioritizes establishing common ground. Taken together, this dissertation provides valuable insights into the cognitive processes that shape our experiences with punishment and highlights the roles they play in facilitating our use of punishment for social control and social pedagogy.

Table of Contents

TABLE OF CONTENTS

List of Figures

LIST OF FIGURES

Acknowledgments

At sixteen years old, I decided to pursue a PhD in psychology. Now, fifteen years later, as I stand on the precipice of achieving that dream, I find myself reflecting on the long winding road that has led me here. Along the way, I have been fortunate to have the friendship, affection, support and guidance of many warm and wonderful people whose presence has shaped not only the journey and its destination but also the person I've grown into during this incredible adventure.

I want to start by thanking my extraordinary mentor, Fiery Cushman. Fiery, the moment you opened the doors of your lab to me, my life changed in ways I could not have imagined. Your mentorship has been nothing short of exceptional, teaching me what it means to be a careful and thoughtful scientist and an integral part of academia, not through mere words, but through your own example. I am grateful for the delicate balance you maintained, knowing when to let me venture into the unknown on my own, and when to hold my hand, offering guidance and help. Your willingness to entertain even the vaguest of ideas, and your dedication to working together to transform them into something precious, have left an indelible mark on me.

I want to thank my committee members: Josh Greene, Tomer Ullman, and Rebecca Saxe. Josh, your insightful questions, helpful critique, and unwavering encouragement, and support have been instrumental in helping me grow as a scientist. I treasure our lighthearted exchanges by the water cooler as much as our engaging conversations about scientific. Tomer, I am grateful for the delightful anecdotes, feedback, and insightful suggestions you shared during the early stages of some of these projects. Rebecca, your razor-sharp thinking, and fearlessness to pursue ambitious research ideas have long been an inspiration to me. The opportunity to discuss my work with you is a gift for which I am grateful.

ACKNOWLEDGEMENTS

I am also grateful to David Lagnado for demonstrating that a life in science can be as enjoyable as one could imagine, to Paul Burgess for agreeing to supervise me even when my research interests ventured beyond his expertise, and for encouraging me to apply to graduate school. To Susan Carey, your developmental proseminar during my first semester of grad school had a transformative impact on my approach to science and I am thankful for the way you've shaped my thinking and work. I'm grateful to Mark Ho, for being not only a collaborator and mentor but also a friend. Mark, your ideas resonate through the pages of this dissertation.

I extend my gratitude to the numerous professors and teachers who have guided me throughout my high school and college years, embracing and, at times, nurturing my spirited nature rather than suppressing it. My heartfelt thanks to Cimeran Kher, Lata Vaidyanathan, Malini Khatri, Amita Parshuram, and Anita Mahajan.

I am grateful to my dear late Anil uncle, who was the earliest supporter of my dreams. His love and the insights he shared through countless stories still serve as beacons of guidance as I traverse the course of my life.

I am grateful to the past and present members of the moral psychology research lab who have contributed to and sustained my scientific journey. The knowledge gained from many, either directly or by observation, has been invaluable. Special thanks to Regan Bernhard, Monica Burns, Adam Morris, Indrajeet Patil, Dylan Tweed, and Wouter Kool for their sass, support, and friendship. Thank you also to the many research assistants and lab managers who enabled the work within and beyond this thesis including Matt Cashman, Catherine Holland, Kirsten Brodie, Isobel Munday, Madhulika Shashtry, Sarah Grevy Gotfredsen, Janani Sekar, Charumathi Badrinath, and Carter Allen. Thank you also to all the wonderful and dedicated members of staff at the Psychology Department without whose support all our work would grind to a halt.

ACKNOWLEDGEMENTS

I'm grateful to my many friends within academia for sharing the joys and tribulations of science with me and to my friends outside of academia for keeping me sane and helping me remember that there is a life beyond the realm of science: Brian Leahy, Dan Janini, Brandon Woo, Lara Kirfel, Adam Ruzicka, Corey Cusimano, Hayley Dorfman, Lauren DiNicola, Evan DeFelipis, Josh Cetron, Pranav Swamy, Pranavi Kapoor, Richa Arora, Ayush Wahi, Anushka Samuel.

A special thank you to Shivani Sharma for being my trusted confidant in navigating life's myriad decisions, both big and small. I am so grateful for your enduring presence in my life. To my cherished Cambridge huddle, Anthony Yacovone, Timothy Conklin, Shirley Wang, Ruben Van Genugten, and Adam Bear. Your camaraderie, affection, and banter have enriched both my research and my life in immeasurable ways.

Thank you to my precious extended families (Rehani's, Bhardwaj's, Samuel's) and grandparents, Dev Raj and Savita Rehani, and Narender Kumar and Svarna Kanta Sarin, I am lucky to have grown up under the shelter of your love and wisdom.

My heartfelt appreciation to my exceptional new family, my in-laws, especially my grandparent-in-law, Prem Gupta and my parents-in-law, Shobhna and Brij Gupta, for showering me with love and care as they would their own child. I struck in-law gold with you all.

My deepest gratitude to my parents, Raju and Punnam Sarin. Words will never be enough to convey the depth of my love and gratitude for you both. Thank you for believing in me and for supporting my decisions, even when they were difficult to understand. Thank you for tirelessly striving to give me a life filled with opportunities that you yourselves never had. Without you, I would not be the person I am today.

ACKNOWLEDGEMENTS

My dearest brother, Akhil, my affection for you is greater than the expressions I often muster. I am endlessly thankful for your unyielding support, and I am inspired every day by the remarkable man you have grown to become.

Lastly, I wish to express my deepest, heartfelt thanks to the person who holds my heart, my best friend, and husband – Varun Gupta. You fill my life with love, kindness, and thoughtfulness, and I count my blessings each day that I am the one you chose to love. Your unconditional love and friendship, and your wicked sense of humor, has brought balance to my life, taught me to find humor in myself, and illuminated life's true treasures. I am boundlessly grateful that I get the privilege of sharing this journey of life with you.

This dissertation is dedicated to my family:
you are the wind in my sails and my port in a storm.

लहरों से डर कर नौका पार नहीं होती,
कोशिश करने वालों की कभी हार नहीं होती ।

- Sohan Lal Dwivedi

The boat that qualms the waves, never gets across,
A person who tries and dares, is never at a loss.

0

Introduction

Among the many ways in which people influence the behavior of others, a particularly powerful one is through punishments. Punishments typically involve imposing an unpleasant or harmful experience onto someone to correct their undesirable behavior, such as time-out for a kid who pushes their friend on a playground, or a fine for an adult who litters. People punish for various reasons – to assert control, to signal their values, to gain a competitive advantage, to educate, or sometimes to simply express their disapproval (Brandt et al., 2003; Carlsmith et al., 2002; Fehr & Gächter, 2002; Funk et al., 2014; Jordan & Rand, 2020; Raihani & Bshary, 2019). Yet, despite these differences in proximate motivations, punishment's adaptive rational is clear enough: It gets people to set, or change, their behaviors in ways that make them better cooperators (although see Raihani & Bshary, 2019 for an exception). What is less apparent however, is *how* punishments achieve this goal. In other words, what are the precise mechanisms by which humans teach with and learn from punishments?

In this dissertation, I examine the cognitive capacities that are crucial to achieve successful punishment between humans as well as the roles they play in shaping our responses to punishment. Specifically, I ask how we conceptualize and structure punishments for teaching

purposes, how we learn from punishments, and how punishments alter our subsequent thoughts and behaviors.

The challenge here is analogous to one faced by linguists in explaining why humans possess a particular set of languages. While the goal of all languages is to facilitate communication and knowledge transfer between individuals, the possibilities for language systems are infinite. Yet, humans developed a finite set of languages that share common syllables and sounds, in part due to the constraints of our vocal cords and auditory systems (Chomsky, 1980; Liberman & Mattingly, 1985; Pinker, 2003). In a similar vein, I want to understand why punishments between humans take the forms that they do by examining the cognitive capacities and constrains that shape how we engage with punishments[1].

With this goal in mind, I consider two key cognitive processes that explain how we perceive and learn from punishments. The first suggests that we treat punishment purely as reinforcements and learn from them by relying on our capacity for reinforcement learning. The second suggests that we treat punishments as a form of communication and learn from them by engaging our capacity for theory of mind reasoning (Baron-Cohen, 2012; Cushman et al., 2019; Ho et al., 2021; Wellman, 2017). I review both in the sections below.

0.1 Punishment as reinforcement

The simplest way for us to learn from punishments is by treating them as a form or reinforcement. To treat something as a reinforcement means to alter our behavior in response to the consequences that follow it. Accordingly, favorable consequences make behaviors more likely to occur and unfavorable consequences make behaviors less likely to occur. Humans display an innate tendency for learning from reinforcements. We readily repeat behaviors that

[1] I focus is on direct, second-party punishments.

have been rewarded and avoid behaviors that have been punished (Skinner, 1938; Thorndike, 1927, 1998) and further exert control over our environments to adapt our behaviors to maximize (actual and expected) rewards and minimize (actual and expected) punishments.

These conceptual and behavioral insights are formalized within the framework of reinforcement learning. The framework offers a precise computational model that describes the various ways people respond to rewards and punishments to learn behaviors that are well suited to their environments (Sutton & Barto, 1998). In the case of punishment, the framework suggests that, on getting punished for an action, people learn to avoid it in one of two ways: Either by updating their representation of the causal contingencies between the action and its preceding states or by adjusting local state-action associations (Botvinick & Weinstein, 2014; Dayan & Niv, 2008; Gläscher et al., 2010). So, if a person gets fined for littering on the road, they may learn to avoid littering in the future either by recognizing that littering causes negative consequences (a fine) or by developing an association between the action of 'littering' and the consequence of a 'fine' within the context of 'being on the road' (where the generalization could be narrower: littering is bad within the context of being on the road with an empty wrapper in hand and no dustbin in sight or it could be wider: littering is bad in general). The exact approach a person uses will have implications for how quickly they learn and how widely they generalize that learning. Yet, both approaches will modify the person's behavior by reducing the likelihood of them littering in the future.

Extensive research, across various tasks and punishment types, demonstrates the efficacy of this simple principle in human learning. Whether it is economic games or parking violations, electric shocks or monetary fines, people are quick at modifying their behaviors in response to getting punished, or in anticipation of it (Azrin & Holz, 1966; Chambliss, 1966; Fehr & Gächter,

2002; Kubanek et al., 2015; Thorndike, 1932). Moreover, people are particularly adept at learning which behaviors to avoid when punished (Azrin & Holz, 1966; Guitart-Masip et al., 2012). This feature of our psychology makes punishment an especially powerful tool to shape the behavior of others. Accordingly, to teach with punishments, a person simply needs to consistently pair a negative consequence with an undesirable behavior or threaten to do so. Overtime, this would prompt a favorable change in the recipient's behavior. There are umpteen examples of people teaching exactly in this manner by pairing negative consequences such as time-outs, fines, and suspensions with undesirable behaviors like bad playground play, littering, and carelessness at work, respectively.

Construing punishments as reinforcements captures the intuitive ways in which we think about and engage with punishments. For instance, many theories on the psychology of punishment conceptualize it as the whip that people crack whenever they encounter undesirable behaviors (Balafoutas et al., 2016; Dugar, 2010; Hofmann et al., 2018; Laurent et al., 2014; Yamagishi et al., 2009). This view also aligns easily with some of the proximate motives of punishment, such as the desire for retribution and offers a nuanced explanation of how punishment can overtime achieve the ultimate goal of deterrence. Thus, thinking of punishments as reinforcements, goes a long way towards explaining cognitive processes that underlie our ability to learn from and teach with punishments. Its simplicity and intuitive appeal make it a valuable tool for understanding the nature of punishment and its role in shaping human behavior.

The explanatory scope of this view however is restricted when it comes to fully explaining punishments between humans. Various experimental findings show that people's natural way of giving out with punishments deviates from the theoretical construct where punishment serves solely as a means of reinforcement. When people are asked to use punishments and rewards to

teach a naïve agent a new task, they structure their feedback in a way that leads a pure reward-maximizing agent astray, sometimes even strengthening the very action they want the agent to avoid (Ho et al., 2018). Similarly, when using punishments to respond to violators of norms, only imposing costly consequences on transgressors does not seem to satisfy punishers. Rather, what seems to matter is whether the transgressor has understood the reason for the punishment (Funk et al., 2014; Gollwitzer et al., 2011; Gollwitzer & Denzler, 2009). In fact, punishers will sometimes willingly impose less costly punishments in favor of more clearly communicating the message behind their punishment (Molnar et al., 2022). These finding are challenging to explain on a view that rests the effectiveness of punishment solely on its direct costliness.

People also don't seem to learn from punishments as if they were just a form of reinforcements. According to the 'punishment as reinforcement' view, the effectiveness of punishment lies in its direct aversiveness. This view it makes no distinction between various sources of punishments, treating punishments generated by the environment (e.g., the sting of a nettle) to be the same as punishments generated by another person (e.g., the sting of criticism). Human learners however are very sensitive to the source of punishment and learn differently based on who has punished them and why. For instance, children appear to draw diametrically opposite inferences of their competence from identical praise or critique given by a teacher, depending on whether or not the teacher had prior knowledge of their abilities (Barker & Graham, 1987; Meyer, 1982, 1992; A. T. Miller & Hom, 1997). Similarly, in instances where people get punished for violating norms, their subsequent response to the punishment, of either cooperating or retaliating, depends on their inference of the legitimacy of the punishment (Baldassarri & Grossman, 2011; Barrett & Harrison, 2005; Bowles & Gintis, 2013; Ertan et al., 2009; Gross et al., 2016; Tyler, 2021; Villatoro et al., 2014). If the person considers the

punishment to be legitimate and the motives of the punisher to be fair, they are inclined to respond positively to the punishment and cooperate in the future. In contrast, if they deem the punishment to be unjustified, they are more likely to retaliate in response.

Together these findings suggest that punishments between humans involves more than just imposing negative consequences and learning to avoid them. Instead, punishers seem to also care about expressing their desired message through the punishment. In turn, recipients and observers pay attention to who has punished them and try to infer why, using the inferences to guide their learning. Punishments between humans take place within intricate social environments, unlike punishments from the environment. When a person receives punishment from the environment, such as the sting of a nettle, their action (touching the nettle) directly produces the harmful outcome (a painful sting). However, when it comes to humans, punishment (timeout) for an action (pushing another kid on the playground) is mediated by another individual (a parent). Thus, punishments between humans occur in rich social contexts, and how a recipient learns from the punishment depends not only on the severity of the punishment, but also on the recipient's inferences about the punisher's intentions and motivations. Consequently, to teach effectively with punishments, people need to carefully select punishment that foster inferences of fairness and legitimacy, instead of simply meting out penalties. Teaching and learning in this manner require the ability to represent and reason about others' minds, and this hinges on our capacity for theory of mind reasoning.

0.2 Punishment as communication

To successfully navigate interactions in our complex social world, we need to understand the actions of the people around us. We do so by relying on our innate ability to explain people's behaviors on the basis of their mental states, such as their beliefs, desires, and knowledge

(Dennett, 1987; Frith & Frith, 2005; Gopnik & Meltzoff, 1997; Perner, 1991; Wellman, 2017). The cognitive ability to do this is known as theory of mind. It's an essential and early emerging aspect of our social cognition, and it allows us to easily comprehend the behaviors of others (Baldwin, 1993; Bloom, 2002). With theory of mind, we can draw rich interpretations from sparse behaviors and make accurate predictions about what actions may follow. For example, we can understand that a person walking towards a food plaza during lunchtime is likely hungry, on their way to get lunch, and will come back with food.

This cognitive capacity offers us an alternative perspective on how individuals engage with punishments, by shifting the focus from the aversiveness of the punitive act to the reasoning behind the actions of both the punisher and recipient. Rather than viewing punishment as a mere imposition of cost, it suggests that punishment is an intentional act in which a punisher uses costly impositions to communicate their message. The message could be anything – from simply expressing disapproval, to signaling one's values, or even just providing the recipient with helpful informative signals to teach them the right course of action. Conversely, the recipient of the punishment can learn by accurately inferring the punisher's intended message. On this view, what makes punishment effective is the inferences expressed and inferred by the punisher and the recipient, respectively. This view does not disregard the importance of the aversiveness of the punitive act. Rather, it proposes that people's response to punishment may depend on the reasons for the punishment and the social context within which the punishment occurs. Costliness of the punishment imposed may be a factor in that consideration, but it is ultimately the social reasoning that drives the process. To engage with punishments in this manner suggests treating it as a form of communication.

Bayesian inference offers a formal probabilistic framework for understanding how we reason about the mental states of others (Baker et al., 2011; Bonawitz et al., 2011; Gweon, 2021; Jara-Ettinger et al., 2016). One of the key tenets of this framework is the notion of inverting one's own generative model to draw rich social inferences from people's overt behaviors (Baker et al., 2009). A generative model is a set of rules that govern how the world works and can be used to generate predictions about what actions or outcomes we could expect to observe in given situations. For instance, if someone desires to quench their thirst and believes that the liquid in the cup before them is water, then, our generative model will predict that the person will reach for the cup and drink the liquid in it.

In the context of theory of mind reasoning, it suggests that people make sense of other people's behaviors by inverting their own generative model to infer a person's covert mental states (desire for water) from their overt behaviors (reaching for the cup). People do so, the framework posits, by engaging in rational belief updating through the use of Bayes' theorem. Bayesian inference assumes that people represent their beliefs about others' mental states, such as their partner's preference for ice cream, as probability distributions rather than as binary choices (likes or dislikes). Then, whenever they encounter new information relevant to that belief, such as seeing their partner pick ice-cream rather than chocolate pudding for dessert, they update their prior belief in light of the new information in a manner consistent with Bayes' rule. Bayesian inference thus provides us with a formal and precise way to understand how people make inferences about other people's mental states from their overt behaviors. In doing so, it provides us with a principled way to understand how people might learn from and teach with punishments.

To learn from punishments successfully a person will need to accurately infer the message their punisher is trying to convey and adjust their behavior appropriately. This involves updating their beliefs about the punisher and the situation and modifying their behavior to avoid future punishment. For instance, if someone is punished by their partner for leaving dirty dishes in the kitchen sink, they use the punishment as a signal to update their understanding of their partner's values and expectations (regarding a clean kitchen) in a manner that approximates precise Bayesian inference. They then use the updated inference to modify their behavior (starting to clean their dishes), to avoid getting punished by their partner in the future. Similarly, when a person, trying to learn a new behavior or gain new information, is punished by a knowledgeable teacher, they can understand the punishment as a signal that their actions or ideas do not align with what their teacher wants them to learn. By abandoning these actions and trying something different, the person can avoid future punishment and improve their understanding and behavior.

To teach effectively with punishments a person will need to craft sanctions and select the punitive actions that will clearly communicate their desired message. This can be done in one of two ways. On the simpler approach, a punisher can select their punishment by referencing their own generative model of a reward-based learner to estimate how the sanction will likely modify the recipient’s behavior, ultimately selecting a punishment that is most likely to achieve the desired change. Punishing in this way will involve one level of mental-state inference according to which the punisher represents the recipient. On a different, more complicated approach, the punisher uses their generative model to select a punishment but then infers what the recipient is likely to infer from the punishment about the punisher and their goals. This inference may then prompt the punisher to modify their chosen punishment, ultimately selecting a punishment that will encourage the recipient to have the beliefs that they want the recipient to hold. Punishing in

this way involves many recursive levels of mental state inference according to which the punisher represents the recipient who represents the punisher representing the recipient. Although reasoning in this way is computationally expensive, compelling research suggests that many instances of pedagogical interactions between humans can be modelled this way[2] (Pinker et al., 2008).

By understanding punishments as a form of communication, specified within the framework of Bayesian inference, we can gain valuable insights into how people learn from and teach with punishments. This perspective helps to explain why people may prefer punishments that are less severe but more informative, as well as why they may choose not to punish in ambiguous situations (Rai, 2022). Additionally, it sheds light on when and why punishments may sometimes be perceived as unfair, leading to retaliatory and anti-social behavior instead of cooperation. Finally, the framework also gives us a deeper understanding of the effectiveness of sanctions by offering predictions on the types of actions that can be impactful, even when they don't come with a direct cost. For instance, you can effectively punish your non-dish-washing partner by giving them a new kitchen sponge and soap. Though the action itself may not impose any negative consequences, the underlying message of your disapproval can be perceived as a form of punishment and inspire them to change their ways.

0.3 Present work

There are then two ways in which people can learn from punishments – by treating them as reinforcements, or by treating them as communication. These ways of engaging with

[2] In principle, the recursion could extend for infinite levels. However, in practice, it is reasonable to assume that there may be a limit to how many levels of recursion we can effectively process. However, the exact extent and limits of our recursive mentalizing abilities are still an area of ongoing research and debate.

punishments are supported by different cognitive processes which although distinct, often work in tandem (FeldmanHall & Nassar, 2021; Jara-Ettinger, 2019; Jara-Ettinger et al., 2020). A typical punishment has elements of both. When a parent issues a time-out for a child's poor behavior, the punishment not only imposes a negative consequence but also signals the parent's disapproval of the child's actions. The aim of this dissertation then is not to create a strict dichotomy between these cognitive processes, but rather to explore their distinct roles in shaping our experiences with punishments. The goal is to understand when the communicative signal of punishment is sufficient for teaching and learning. When is the aversive aspects of negative reinforcement necessary? How do these processes interact in shaping the dynamics of punishments in everyday human interactions? By delving into these questions, my aim is to shed light on the cognitive capacities and constraints that enable us to effectively use punishment to influence the behavior of others.

In Chapter 1, I investigate how individuals conceptualize and engage with punishments in response to common violations of norms. Specifically, I examine whether people treat punishment solely as a reinforcement or as a mode of communication. I compare the effectiveness of canonical punishments, which are both costly and communicative, to punishments that are costless but communicative, as well as those that are costly but uncommunicative. The findings indicate that people consider costless but communicative punishments to be as effective in changing a transgressor's behavior as canonically costly punishments. These results suggest that people's understanding of punishment is based on principles of communicative inference. As a result, punishers select punishments by reasoning about the recipient's mind, using punishment to communicate their desires set of values and norms. Recipients of the punishment, in turn, learn from them by accurately inferring the

punisher's communicative intent. The findings from Chapter 1 underscore the role of 'punishment as communication' in shaping our everyday punishments.

In Chapters 2 and 3, I explore the effectiveness of punishments in instances where norm violations are committed unwittingly, such as in cases of negligence. Negligence occurs when someone causes harm that they could have foreseen but failed to due to an automatic failure of thought. In such cases, punishing someone presents a challenge for the recipient trying to learn from the punishment as simply inferring the punisher's communicative intent may not be enough to prevent the transgression from happening again. While communication-based punishments can address intentional carelessness, they may not be as effective in cases of automatic failures of thought. However, drawing on insights from cognitive science, I argue that punishment's negative reinforcement will have an automatic impact on the thoughts that come to mind in the future. Thus, in these situations, the aversiveness of the punishment alone can prompt successful changes in future behavior. Consequently, insights from Chapters 2 and 3 emphasize the relative importance of the 'punishment as reinforcement' proposal in certain circumstances.

Finally, in Chapter 4, I focus on the use of punishment in instances of social pedagogy, where a teacher aims to impart knowledge to a learner using evaluative feedback. In two experiments, I investigate how teachers guide the behavior of novice learners through the use of punishments and rewards and propose a computational model to capture their strategy. The findings from Experiment 1 indicate that teachers tend to structure their feedback communicatively to modify the learner's beliefs. The results from Experiment 2 build on this insight by suggesting that when teachers modify a learner's beliefs, they place greater emphasis on establishing a shared understanding between themselves and the learner, using it to facilitate subsequent teaching.

Together, this dissertation provides valuable insight into the cognitive processes that govern our use of punishment for teaching and learning. Through the findings of each chapter, it becomes evident that both the reinforcement of punishment as well as it's communicative signal play important roles in shaping our use of punishment to influence others' behavior. But by understanding the relative contributions of these processes, we can develop a more nuanced and comprehensive model of human punishment, that can serve as a unified framework for future research. Furthermore, a deeper understanding of these processes can help develop more effective and compassionate approaches to disciplining others in ways that promote positive interactions and foster healthy relationships. This research also has important implications for the field of Human-Computer Interaction, as it can help clarify which aspects of the learning process must be explicitly specified for AI agents to learn successfully from feedback provided by humans. Ultimately, the pursuit is valuable because understanding the mechanisms that make punishments effective, will enable us to develop precise scientific theories, more equitable and effective disciplinary measures, and design smarter AI systems.

1

Punishment as communication

Abstract

Humans use punishment to influence each other's behavior. Many current theories presume that this operates as a simple form of incentive[3]. In contrast, we show that people infer the communicative intent behind punishment, which can sometimes diverge sharply from its immediate incentive value. In other words, people respond to punishment not as a reward to be maximized, but as a communicative signal to be interpreted. Specifically, we show that people expect harmless, yet communicative, punishments to be as effective as harmful punishments (Experiment 1). Under some situations, people display a systematic preference for harmless punishments over more canonical, harmful punishments (Experiment 2). People readily seek out and infer the communicative message inherent in a punishment (Experiment 3). And people expect that learning from punishment depends on the ease with which its communicative intent

[3] This chapter originally appeared in the literature as: Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544.

can be inferred (Experiment 4). Taken together, these findings demonstrate that people expect punishment to be constructed and interpreted as a communicative act.

1.1 Introduction

Unruly pets are scolded; disobedient children are put in time out; hardened criminals are imprisoned—when humans encounter bad behavior, they will often punish it. The prototypical punishment imposes a cost on the transgressor, and its primary function is to change the recipient's behavior (Boyd & Richerson, 1992; Clutton-Brock & Parker, 1995; Fehr & Gächter, 2002). But it is less clear exactly how punishment changes behavior. In principle, there are many different ways in which people might learn from punishment, and many corresponding ways in which punishment might be designed to generate this learning.

We focus on two broad possibilities. The first, obvious possibility is that the experience of punishment is designed to exploit simple reward learning. On this view, punishment is a "corrective whip that acts as a negative reinforcer of desired values" (Hofmann et al., 2018). Although humans have sophisticated and sometimes specialized learning mechanisms, punishment may not depend on these in order to work. Rather, it may require learning mechanisms no more complex than Thorndike's Law of Effect (1927): Repeat what has been rewarded; avoid what has been punished. This view of punishment is widely supported, and most psychological theories of punishment commit to it implicitly or explicitly. One of its principal merits is minimalism: by construing punishment as a direct incentive scheme constructed by the punisher, it seems to capture its essence. Parsimoniously, this view construes learning from social punishment as fundamentally similar to learning from non-social experience.

Yet, our everyday experiences belie this: when a person gets whipped by the wind, they avoid wind. But when a person feels the corrective whip of punishment, they do not simply avoid

the punisher. Instead, they try to understand what the punisher meant. Thus, we suggest that at least between humans, punishment may be designed not only as an incentive, but also as a means for the punisher to convey their message of frustration and disapproval to the target. In keeping with this possibility—and building on prior philosophical (Ewing, 1943; Feinberg, 1965) and psychological research (Funk et al., 2014)—we suggest that punishment is designed to be interpreted as a form of communication (Cushman et al., 2019). On this view, transgressions are teachable moments that punishers exploit, using punishment to communicate their expected set of values and norms. Recipients, in turn, interpret the punishments as communicative acts and learn by both responding to the costs as well as by accurately inferring the punisher's intended message. This involves a sort of meeting-of-the-minds between the punisher and the recipient. It requires mental state inference on the part of both parties, as each of them attempts to guess how the other will interpret and anticipate the communicative act. This model therefore proposes that we learn from social punishment very differently than from non-social experiences (like a whipping wind), which are not typically interpreted as intentional communicative acts. Learning from social punishment therefore engages social cognition and inference, much like other forms of communication like language (M. C. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Grice, 1957; Kao et al., 2014; Kao & Goodman, 2015; Sperber & Wilson, 1986) or demonstration (Ho et al., 2016; Shafto et al., 2014).

In summary, we contrast two models of how human punishment is designed. The first is as a "constructed incentive": The goal of the punisher is to set up a simple contingency between bad action and bad consequences, relying on the target's basic capacity for reward learning to adjust in response to this incentive. According to the second model, punishment is both a constructed incentive and a communicative act. The unique feature of this model is that punishment is

designed to be interpreted by a learner who assumes communicative intent on the part of the punisher. Of course, this second model does not deny the incentive value of punishment, but instead augments it. Both models agree that punishers wish to modify the behavior of transgressors, and that transgressors wish to understand the behavioral modification intended by punishers. They disagree on a further question: whether behavioral modification is organized around simple principles of incentive (Dayan & Niv, 2008; Sutton & Barto, 1998), or around more complex principles of inferential communication akin to natural language pragmatics (Frank & Goodman, 2012; Grice, 1957; Sperber & Wilson, 1986). Our goal is to distinguish the key predictions of these models experimentally.

1.1.1 The psychology of punishment

At an ultimate level, one of punishment's functions is to modify the behavior of social partners (Boon et al., 2009; Cushman, 2015; Hampton, 1984; Heider, 1958; Martin & Cushman, 2016; Miller, 2001; Morris, 1981; Smith, 1869; but see Raihani & Bshary, 2019 for alternative functions). It typically occurs when somebody has caused harm or violated a norm, and it typically serves to both modify that person's future behavior and also indirectly influence others who observe it. The mechanisms responsible for accomplishing this function, however, are debated.

One body of evidence contends that punishment is motivated by a relatively simple taste for revenge (Carlsmith, 2006; Carlsmith et al., 2002). On this view, when people are harmed, they experience a direct motivation to cause harm in retaliation (McCullough et al., 2013). Thus, people often punish in situations where doing so does not serve them well rationally (Lerner & Clayton, 2011). For instance, people punish in one-shot interactions (Balafoutas, Grechenig, et al., 2014; Fehr & Gächter, 2002) and even when punishing makes them feel bad (Carlsmith et

al., 2008). Such retributive instincts may serve the adaptive function of modifying others' behavior, but without much role for a proximate psychological motive of behavior modification (again, much like hunger motivates nourishment, but without much role for the proximate psychological motive of nourishment).

To the extent that human punishment amounts to simple and pure retribution, this aligns best with the "constructed incentive" model of punishment. The essential, "eye-for-an-eye" logic of retribution simply involves imposing a cost on others whenever a cost is experienced for oneself. This mechanism neither assumes nor requires that punishments be interpreted as communicative acts.

In contrast, a rich tradition in philosophy argues that a primary function of punishment is to express disapproval of the perpetrator (Durkheim, 2014; Feinberg, 1965; Stephen, 2014) and to educate both the offender and the public (Ewing, 1943). This view, collectively referred to as expressionism (Primoratz, 1989; Skillen, 1980), suggests that the cost imposed on a perpetrator is meaningful because it communicates society's condemnation of the crime. A growing body of experimental research examines this dissertation by focusing on the thoughts and actions of punishers. This work suggests that human punishment is generated by more sophisticated motives than the simplest models of retribution suggest. For instance, punishers are satisfied with their punishments more when the target accurately understands their intent to punish and additionally, signals a change in his/her attitude towards his/her wrongdoing (Funk et al., 2014; Gollwitzer et al., 2011; Gollwitzer & Denzler, 2009). In the absence of such an acknowledgment, punishers are dissatisfied with the punishment no matter how strong of a cost it levies on the target (Funk et al., 2014). In fact, even those punishers who typically punish harshly will willingly punish less severely if that would allow the target to better understand their intent

(Molnar et al., 2022). As these findings demonstrate, the desire to communicate one's beliefs, values, and disapproval plays a central role in the kind of punishment a punisher chooses to enact.

This accords with recent formal work showing that people seem to give our rewards and punishments to communicate and express approval rather than to simply incentivize an action. In one such set of experiments, participants played a computer game in which their job was to teach a child or puppy the right way to navigate to a goal (Ho et al., 2018). Participants could only teach by responding to the learner's behavior with rewards and punishments. Quantitative analysis showed that people had substantial difficulty teaching pure reward-maximizing agents (i.e., agents programmed to learn through the Law of Effect), but easily taught agents who interpreted feedback in terms of participants' communicative goals.

In sum, this body of work suggests that punishers want to be "understood", effectively modifying the behavior of those who they punish. Yet, it remains an open question exactly how they try to achieve this goal. Do they rely on a simple incentive scheme by imposing observable costs on bad behavior? Or do they rely on a more complex cognitive model of the inferences that an observer would draw in order to select specific punishments? Similarly, we know little about how people learn from punishment. Do they respond to the punishments merely as an incentive, or do they model the communicative intent of the punisher? The question at stake is not whether punishment, and learning from it, is designed to communicate something - it clearly is. It is also not whether people have the general capacity to express and infer communicative intent; we know this is essential to natural language, for instance (Frank & Goodman, 2012; Grice, 1957). The question is whether these aspects are connected. Is the "message" communicated by punishment limited to the immediate imposition of material costs? Or is punishment more like

natural language: a form of communication organized around principles of communicative inference? On the latter view, punishers could express, and recipients could accurately infer, the "message" even in the absence of immediate, imposed incentives. This is the key idea motivating our experiments.

1.1.2 Experimental logic

A typical punishment is a mixture of two ingredients: the incentive it imposes and the message it communicates (see Figure 1.1). Our goal is to examine which of these elements most accurately captures the necessary and sufficient conditions that make punishment work. To do so, we need to investigate the parts of this diagram that do not overlap. Do punishments that impose incentives but lack communicative elements work? Conversely, how effective are punishments that impose no cost but carry clear communicative elements?

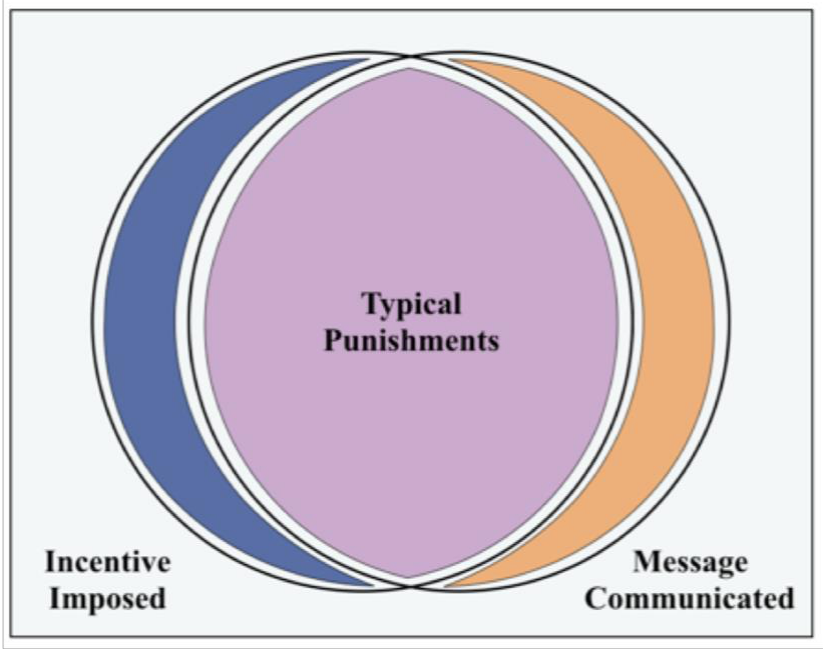

Figure 1.1: Conceptual diagram of experimental logic.

Thus, we explore a setting in which punishment will only be successful if the recipient models the communicative intent of the punisher and cannot succeed based on simple incentive

alone. We accomplish this by constructing cases where the logic of communication allows punishment to work not by imposing a cost on transgressions, but instead by providing an immediate benefit.

How can providing a benefit to somebody play the communicative role of punishment? Our approach is inspired by models of figurative speech. Instances of figurative speech, such as verbal irony, hyperbole, etc. (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Kao et al., 2014; Kao & Goodman, 2015) entail statements whose literal meanings are different from the speaker's intended meaning (Colston & O'Brien, 2000; Roberts & Kreuz, 1994), something listeners seem to easily understand. Existing models of language suggest that listeners interpret statements not only by attending to their literal meaning, but also by inferring their potential figurative meaning from a model of the speaker's communicative intent. Therefore saying "What great weather we have today!" in response to a cold, stormy day literally conveys the speaker's enjoyment of the weather but figuratively and pragmatically communicates their displeasure.

Inspired by this work on figurative speech, we created instances of 'figurative punishment': punishments that could be interpreted appropriately as punishments only by considering the punisher's communicative intent. For example:

Alice lives in an apartment with three other roommates. One of them, Sandra, has a very frustrating habit of leaving her dirty dishes in the kitchen sink. The roommates have pointed out the dishes to Sandra, but they don't know exactly how to say the right thing, and she keeps doing it. Alice talks with her other roommates and together they decide to do something about it. The next day, before Sandra gets back home, Alice leaves a brand-new kitchen sponge and bottle of dish soap on her pillow with a ribbon around them and a tag that says: "Love, your roommates".

Alice's gift to Sandra imposes no cost. On a model where recipients treat punishment as an action-incentive contingency to be maximized, Sandra should not change her behavior in light of the sponge and dish soap. If anything, she should double-down because the items impart a slight benefit. And yet, understood pragmatically, in light of her roommate's communicative intent, the sponge and dish soap feel like sanctions and lead Sandra to correct her behavior. This is because (as we show experimentally) the most plausible interpretation of Alice's behavior is that it communicates frustration with Sandra's failure to do the dishes, not approval of it. Of course, this presumably motivates Sandra because it implies that she will eventually pay costs for not doing dishes. Alice's action may therefore be effective because it communicates a threat. It may also engender feelings of guilt. However, the threat, the potential for guilt, and any associated social and emotional costs experienced stem from the inferred communicative intent of Alice's act and not from the act itself. In other words, the force motivating Sandra to change her behavior does not follow directly from the value of Alice's act, as the constructed incentive model demands. Instead, it follows from the inferred communicative intent of Alice's act, as our model allows.

We study these cases of figurative punishments not because they are normal, but because they provide us with a useful case study. Studying figurative punishments can shed light on the psychological underpinnings of typical "literal" punishment, much in the same way that studying figurative speech, such as irony and hyperbole, can inform our understanding of regular speech. According to our account, even ordinary punishments may be interpreted in light of their apparent communicative intent. But it is hard to know because a simple constructed incentive model also predicts that we can learn from these ordinary punishments. Studying figurative

punishment is useful because, despite being uncommon, it is highly diagnostic of the underlying psychological mechanisms.

It is worth noting that our focus in this research is on acts and whether they are interpreted in light of the actor's communicative intent, not on speech. Even though we propose that interpreting social punishments is cognitively akin to understanding language, actions differ from speech in many ways (see General Discussion for a more in-depth discussion on this). One fundamental difference is that language relies on conventionalized semantics that relate meanings to words, whereas arbitrary actions do not have conventionalized meaning. Additionally, while much work investigates linguistic pragmatics, less work in social and cognitive science has examined the principles that ground inferences about communicative actions. Understanding this, within the context of punishments, is the focus of our research.

Our studies, all pre-registered[4,5] use several complementary methods to identify whether punishment is designed to impose immediate, direct costs on social partners who transgress, or instead to provide informative communicative signals. Experiment 1 asked whether people think that figurative punishments can be informative and effective. Experiments 2a and 2b asked whether people ever prefer figurative punishments to literal ones and how context influences their choice. Experiment 3 asked whether people infer the communicative intent of figurative punishments even in ambiguous cases. Finally, Experiment 4 asks whether "literal" punishments are less effective in the absence of clear communicative intent. We report on all the dependent

[4] Pre-registered documents can be found at the following links: Experiment 1 (https://aspredicted.org/9g5zj.pdf); Experiment 2a (https://aspredicted.org/ 8cn5k.pdf); Experiment 3 (https://aspredicted.org/vv5t3.pdf); Experiment 4 (https://aspredicted.org/7de6p.pdf)

[5] Sample size for each study was calculated based on pilot data and power analyses to obtain at least 80% power of detecting the reported effects.

variables that were collected for the experiments. Data and analysis scripts for all experiments can be found on OSF at https://osf.io/fxdwm/.

1.2 Experiment 1

We begin by testing whether people believe that a "figurative punishment" could ever possibly be an effective way of changing someone's behavior. In other words, do they believe that an immediate, direct "incentive" is necessary to make punishment work, or do they believe that a communicative act alone can be sufficient? We described to participants several hypothetical cases of "literal" versus "figurative" punishments. We designed both types of punishment to have clear communicative intent to express displeasure with a transgression (e.g., leaving dirty dishes piled up in a communal sink). What differed was the actual, immediate consequences of the punishment: Literal punishments imposed a cost (e.g., dirty dishes left on a roommate's pillow), while figurative punishments imposed a slight benefit (e.g., a new sponge and bottle of dishwashing detergent left on the roommate's pillow).

In two separate conditions we also tested whether people believe that clear communicative intent is necessary for punishment to work, or whether they instead consider an imposed cost (i.e., incentive) to be necessary. To do this we constructed baseline conditions that were matched in the relative costs and benefits of the literal and figurative punishments, respectively, but were unclear in their communicative intent (that is, difficult to interpret given the context). For instance, in a case where the roommate has been leaving her dishes undone, the punisher either throws her mail into her room in a disorderly manner, or else stacks it very neatly in her room. These behaviors impose immediate costs and benefits, respectively. But crucially, they have no transparent connection to dishes. Therefore, their communicative intent is obscure. We asked people whether behaviors like these would-be effective forms of punishment.

Thus, we have a 2 × 2 design: punishments either imposed a cost or were costless and they either clearly communicated the punisher's intent or not. This yielded the following four conditions: Literal (costly) and figurative (costless) punishments with transparent communicative intent, and matched non-communicative versions of each (costly and costless baselines). Participants read one of the four possible kinds of punishment and answered a series of questions meant to assess their expectations about the effectiveness of the punishment as third-party observers. If people believe that imposing a cost is sufficient to make punishment effective, then the two conditions imposing costs (literal punishments and baseline costly punishments) should be judged effective. If they believe that clear communicative intent is sufficient to make punishment effective, then the two conditions with clear communicative intent (literal and figurative punishments) should be judged effective. It is also possible that people will judge both of these factors sufficient—in this case, the only ineffective condition would be the baseline costless condition. Or it is possible that they will judge both factors necessary, in which only literal punishments will be judged effective. In addition to Experiment 1 that tested participants' expectations as uninvolved third-party observers, in two separate experiments we also manipulated perspective-Experiment 5 (Appendix A) put the participants in the role of the punisher and Experiment 6 (Appendix A) put them in the role of the recipient. Results were qualitatively similar across changes in perspective. As a result, we focus on (third-party) observer perspective results below and report results from Experiments 5 & 6 in Appendix A.

1.2.1 Method

Participants 2000 people were recruited using Amazon's Mechanical Turk. We retained 1953 participants after excluding those who failed our attention checks (2.4% excluded). 56% of the sample was female and the modal age range was 31–40 years. Participants were paid $0.33.

Design & materials Eight everyday scenarios were created. In each scenario, one agent committed a norm violation (for instance, leaving her dirty dishes about) and another agent responded to the violation by punishing the transgressor (for instance, by stacking the dirty dishes on her pillow). The punishment always involved leaving some "item" for the transgressor. The item was either beneficial for the transgressor or costly (thus, "figurative" or "literal") and was either contextually related to their transgression or not (thus, "informative" or "uninformative"), leading to a total set of $8 \times 2 \times 2 = 32$ vignettes (see A.1 in Appendix A for the list of scenarios used). Participants were randomly assigned to read one scenario.

In order to develop these stimuli, we created a set of candidate "items" (e.g., dirty dishes, or a brand-new sponge and dish soap) and pretested these on an independent sample of participants ($N = 103$). They rated how they would feel upon receiving each item (from terrible (0) to delighted (10)). We did not provide any context regarding a transgression—thus, these participants were unlikely to interpret any of the items as a form of punishment. Only items with mean negative (<5) or mean positive ratings (>5) were used in the final vignettes. To further control for variance caused by the items themselves, each pair of items was used twice: once in an "informative" condition (where they were semantically related to the transgression) and once again in an "uninformative" condition for another vignette context where they were semantically unrelated to the transgression (see A.1, Appendix A).

Following the presentation of a single vignette, all participants answered a series of questions in the order presented here. Participants first rated the interpretability of the punishment: "Will Sandra get the message that she needs to start cleaning her dirty dishes?" (0 = definitely not; 10 = definitely yes). Next, they judged the effectiveness of the punishment: "How likely is Sandra to start cleaning her dirty dishes?" (0 = very unlikely; 10 = very likely).

Following this, they rated their perception of how the recipient of the punishment would feel: "When Sandra sees what her roommates have left for her, how will she feel?" (0 = terrible; 10 = delighted). Finally, they judged if the punishment would feel like a punishment to the recipient: "Would this feel like a punishment to Sandra? " The last rating was made on a binary scale with labels 'definitely yes' and 'definitely no'. Each scenario included one comprehension question and participants failing to answer this question correctly were excluded from all analyses.

Statistical methods

To analyze participant responses, we fit three linear mixed effects models and one generalized linear fixed effects model (Gelman & Hill, 2006; Jaeger, 2008) using the lme4 package in the R statistical computing environment (Bates et al., 2015; R Core Team, 2022).

All the models were run with effect-codes for cost imposed (costless = 0.5, costly = 0.5), informativeness of the punishment (uninformative = 0.5, informative = 0.5) and their interactions. Following our preregistered analysis, the analysis reported below has a random intercept only for the stimuli (DV ~ cost imposed * informativeness + (1| item), data = data). In the Appendix A (A.6), we present additional analysis with a full random effect structure (Barr et al., 2013) along with our rationale for choosing the analysis presented below (Bates et al., 2015). The results from both sets of analysis (the one presented below and the maximal random effect model) converge on all qualitative findings.

1.2.2 Results

We first looked at participants' expectations about each punishment and how clearly it would allow the recipient to 'get the message' (Figure 2.1 (a)). We found a main effect of cost imposed, $b = 0.50$, $t_{1942.03} = 3.90$, $p < .001$, a main effect of informativeness, $b = 3.77$, $t_{1942.17} = 29.53$, $p < .001$, and a significant cost × informativeness interaction, $b = 0.95$, $t_{1942.15} = 3.72$, $p <$

.001. To interpret this predicted interaction, we carried out planned pairwise contrasts using the emmeans package (Lenth et al., 2018). When the punishment was informative and communicative, the cost it imposed did not matter, as participants rated the recipient to be equally likely to get the message from costly, literal punishment (e.g., dirty dishes; $M = 5.87$, $SE = 0.20$, CI: [5.44, 6.31]) as from a costless, figurative punishment (e.g., new sponge and dish-soap; $M = 5.85$, $SE = 0.20$, CI: [5.42, 6.28]), $t(1942) = 0.13$, $p = .90$. However, when the punishment was uninformative, participants judged the recipient to be more likely to get the message from a costly punishment (e.g., mail thrown around messily; $M = 2.58$, $SE = 0.20$, CI: [2.15, 3.01]), compared with a costless punishment (e.g., mail stacked away neatly; $M = 1.61$, $SE = 0.20$, CI: [1.17, 2.04]), $t(1942) = 5.38$, $p < .001$.

Next, we looked at the perceived effectiveness of each punishment in bringing about a change in the recipient's future behavior (Figure 1.2 (b)). The results revealed a main effect of cost, $b = 0.55$, $t_{1942.11} = 4.71$, $p < .001$, of informativeness, $b = 2.53$, $t_{1942.51} = 21.49$, $p < .001$, and a significant interaction between the two, $b = 0.75$, $t_{1942.44} = 3.16$, $p = .002$. Planned pairwise contrasts showed that when punishments were informative, costless punishments ($M = 4.64$, $SE = 0.14$, CI: [4.36, 4.92]) were perceived to be as effective in changing behavior as costly punishment ($M = 4.82$, $SE = 0.14$, CI: [4.54, 5.10]), $t(1942) = 1.09$, $p = .27$. However, when punishments were uninformative, the cost imposed made a difference as costly punishments ($M = 2.66$, $SE = 0.14$, CI: [2.39, 2.94]) were perceived to be more effective in changing behavior than costless punishments ($M = 1.74$, $SE = 0.14$, CI: [1.46, 2.02]), $t(1942) = 5.56$, $p < .001$.

Participants' judgments regarding the recipient's feelings upon receiving the punishments also showed a main effect of cost imposed, $b = 2.45$, $t_{1942.03} = 24.47$, $p < .001$, of informativeness, $b = 0.46$, $t_{1942.14} = 4.60$, $p < .001$, and a significant interaction, $b = 1.00$, $t_{1942.12} =$

4.97, $p < .001$ (Figure 2 (c)). Planned pairwise contrasts revealed that, when the punishment imposed a cost, participants expected the recipient to feel bad, irrespective of whether the punishment was informative ($M = 2.09$, $SE = 0.17$, CI: [1.71, 2.46]) or uninformative ($M = 2.05$, $SE = 0.17$, CI: [1.68, 2.42]), $t(1942) = 0.26$, $p = .80$. However, when the punishment was costless, participants judged the recipient to feel significantly worse (as evident by the lower mean score) when the punishment was informative and directly related to the transgression ($M = 4.04$, $SE = 0.17$, CI: [3.67, 4.41]) than when it was uninformative and unrelated to the transgression ($M = 5.00$, SE = 0.17, CI: [4.63, 5.37]), $t(1942) = 6.75$, $p < .001$.

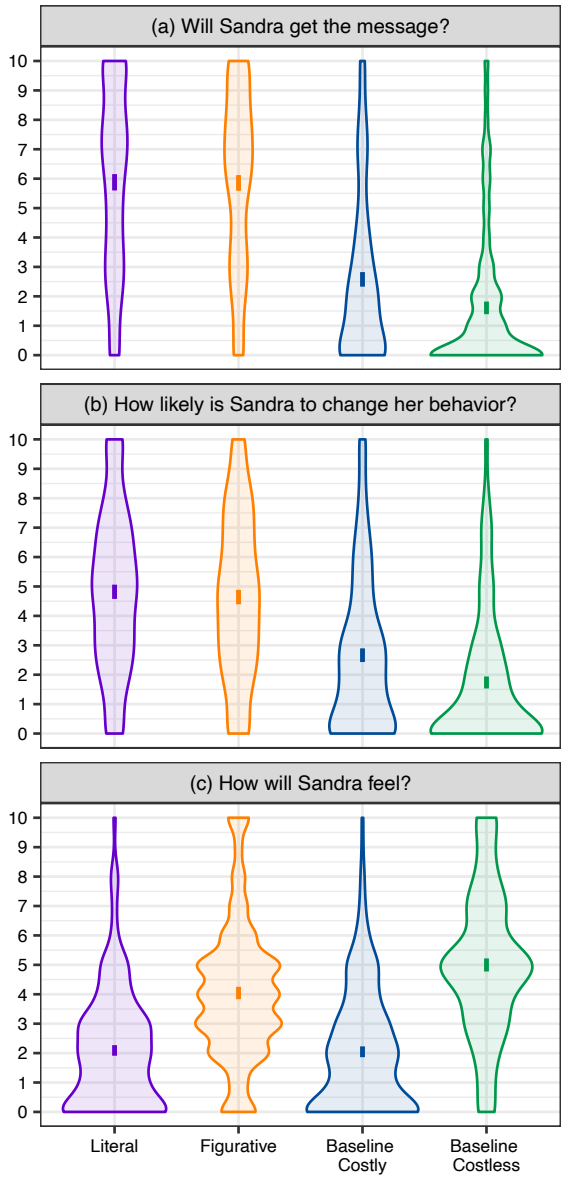

Figure 1.2: Distribution of participant responses for three of the four dependent variables in Experiment 1. Lines inside the violins represent 95% CIs of means.

Finally, we examined ratings of whether the act in question would feel like a punishment (results are on the log odds ratio scale). We found a main effect of cost imposed, $b = 1.48$, $SE = 0.10$, $z = 14.70$, $p < .001$, informativeness, $b = 0.91$, $SE = 0.10$, $z = 9.05$, $p < .001$, and a significant interaction, $b = 0.50$, $SE = 0.20$, $z = 2.45$, $p = .01$ (Figure 1.3). Planned contrast revealed that, when the punishment was costly, informative punishments ($M = 1.08$, $SE = 0.15$, *CI*: [0.79, 1.36]) were judged to feel more like punishments than uninformative punishments (M

= 0.41, *SE* = 0.14, *CI*: [0.14, 0.68]), $z = 4.76$, $p < .001$. A similar pattern of results was found for costless punishments, as informative punishments were judged to feel more like punishments (*M* = 0.16, *SE* = 0.14, *CI*: [0.42, 0.11]) than uninformative punishments (*M* = 1.32, *SE* = 0.15, *CI*: [1.61, 1.02]), $z = 8.00$, $p < .001$.

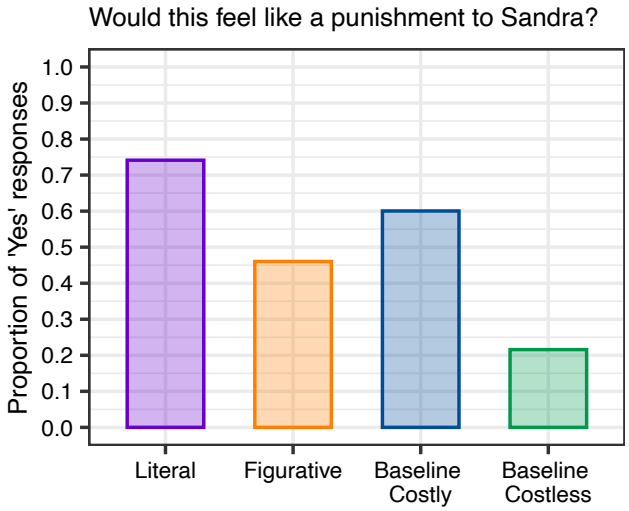

Figure 1.3: Proportion of participants who think the action would feel like a punishment to the target.

1.2.3 Discussion

What do people think is required in order for a punishment to be effective? We find that people think a punishment must be "informative"—i.e., sufficiently semantically related to the offense that its communicative intent is apparent. In contrast, we find that people do not think the punishment needs to impose a cost. Rather, participants in our experiment expected costless, benefit-imposing (figurative) punishment to bring about future behavioral change in a recipient, and in themselves when they are described as the recipients (Experiment 6 Appendix A). Although figurative punishments do not impose a literal cost, people think that the recipients of (informative) figurative punishments, including themselves, will feel bad, presumably because they will understand the underlying communicative intent. The same costless, beneficial

punishments, however, were not judged to be effective sanctions if they were not semantically related to the transgression, and therefore uninformative. These inferences remain stable across changes in perspective – participants judge figurative punishments to be effective punishments as uninvolved third-party observers (Experiment 1), as punishers themselves (Experiment 5 Appendix A) and even as recipients of these punishments (Experiment 6 Appendix A). Taken together, these results suggest that the psychology underlying punishments is better understood as organized around principles of effective communication rather than principles of reinforcement-based learning.

While Experiment 1 investigates whether people can make accurate inferences from costless, figurative punishments, it does not tell us what kinds of punishment people would prefer to actually employ. This is our focus in Experiment 2.

1.3 Experiment 2

In Experiment 2a and Experiment 2b we investigated the kinds of punishments people voluntarily select and the role context plays in their selection.

1.3.1 Experiment 2a

We presented participants with the same set of vignettes and candidate punishments used in Experiment 1. We then asked them to indicate which punishment they would choose to perform, presenting them with four candidates in a 2 (costly vs. costless) × 2 (informative vs. uninformative) design. Our main goal was to see whether participants would select figurative punishments in the presence of literal alternatives. For instance, we wanted to see whether in Alice's position, they might prefer to leave a clean sponge and dish soap on Sandra's pillow rather than leave Sandra's dirty dishes on her pillow.

1.3.1.1 Method

Participants Based on a power analysis reported in our preregistration, 88 participants were recruited using Amazon's Mechanical Turk and paid $0.25 for their participation. After excluding those who failed to pass the comprehension check, the final sample was made up of 81 participants (8.0% excluded; 56% female, modal age range: 31–40 years).

Design & materials The eight base scenarios used in Experiment 1 were employed. As in Experiment 1, each scenario included a norm violating agent. Unlike Experiment 1, however, participants were asked to imagine that they were on the receiving end of the transgression and were asked how they would respond to the norm-violating agent. Participants were presented with four potential responses that corresponded to the four kinds of punishments used in Experiment 1: literal, figurative, baseline costly, and baseline costless. They were asked to select one of the four and then give reasons for their selection. Each participant was randomly assigned to read one vignette.

1.3.1.2 Results

Collapsing across different contexts, a chi-square test of goodness-of fit was performed to determine whether the four kinds of punishments were preferred equally. We found evidence in favor of the alternative hypothesis $X^2(3) = 114.51$, $p < .001$, $V_{Cramer} = 0.69$, *CI*: [0.58, 0.81], n = 81. Descriptive analysis revealed that overall, figurative punishments were selected most frequently (75%, *CI*: [67%, 84%]), followed by literal punishments (19%, *CI*: [10%, 27%]), baseline costly punishments (4%, *CI*: [0%, 13%]) and baseline costless punishments (2%, *CI*: [0%, 11%]) (Figure 1.4).

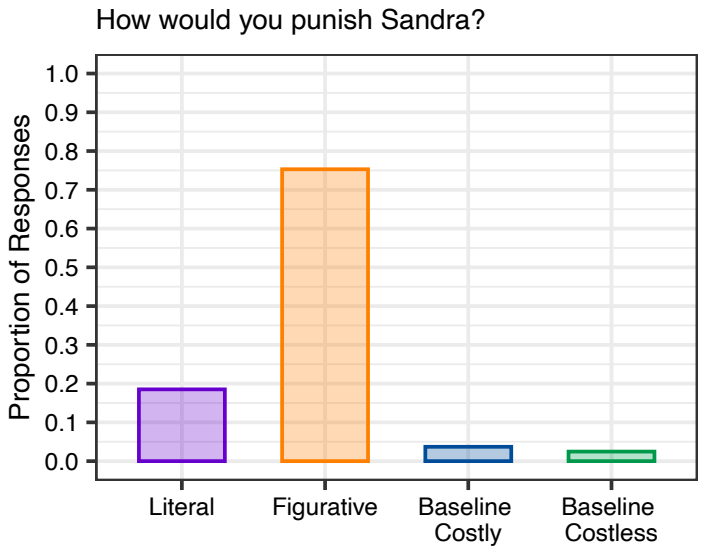

Figure 1.4: Proportions reflecting which punishment participants would choose themselves as punishers.

This preference ordering held in every context (except for one, see Table 1.1). Chi-square tests conducted simultaneously within each scenario, using the ggstatsplot package (Patil, 2021), provided evidence for the alternative hypothesis and a preference for the figurative punishment for all but one context (Table 1.1).

Table 1.1

Choice decision broken down by scenario.

		Percent selected						
Context	N	Literal	Figurative	Baseline Costly	Baseline Costless	Chi-square statistic	p.value	df
Sweaty clothes	8	12.50%	75%	12.50%	0	11	1.17e-2	3
Stationary	9	22.22%	66.67%	0	11.11%	9.22	2.65e-2	3
Loud music	12	8.33%	83.33%	0	8.33%	22	6.52e-5	3
Messy mail	9	0	100%	0	0	27	5.89e-6	3
Laundry	12	8.33%	83.33%	8.33%	0	22	6.52e-6	3
Hair	11	18.18%	72.73%	9.09%	0	14.1	2.78e-3	3
Fridge food	10	20.00%	80%	0	0	17.2	6.43e-4	3
Dirty dishes	10	60%	40%	0	0	10.8	1.29e-2	3

1.3.2 Experiment 2b

Experiment 2a revealed that people prefer costless yet communicative punishments to costly, typical ones. This preference was consistent across the various scenarios we used. Yet, the only decision people could make in Experiment 2a was the kind of punishment they could impose. In contrast, in real life a frequent first step towards confronting norm violating behavior is to talk directly with the transgressor. Moreover, though people displayed a consistent preference for figurative punishments in the cases we employ, we nevertheless suspected that some features of a context (punisher-transgressor relationship, frequency of the offense, etc.) would influence when figurative punishments are preferred over literal ones. In Experiment 2b we addressed both of these concerns. We manipulated the frequency of previously communicating the offense to see whether this impacted the response our participants chose. We also included the option to 'talk' to the perpetrator. In two additional experiments reported in

Appendix A, we tested whether the nature of relationship between the punisher and transgressor (friend vs strangers in Experiment 7a and landlord vs tenant in Experiment 7b) affected which punishment is selected.

1.3.2.1 Method

Participants We recruited 161 participants on Amazon's Mechanical Turk in exchange for $0.25 for their participation. We excluded those who failed our attention check, leaving 158 participants in the final sample (1.9% excluded; 57% female, modal age range: 31–40 years).

Design & materials We modified one of the vignettes used in Experiment 2a. Participants read about the norm-violating agent Sandra, who had not been doing her dishes. Then in a between-subjects design they read that Sandra's roommates had either asked her to clean up on three previous occasions or had never brought up the dirty dishes with her before. In other words, they had either responded to the norm-violating behavior many times, or this was their first time addressing it. After reading the vignette we asked our participants how they would respond to Sandra. We gave them a choice between talking to her one-on-one, leaving a brand-new dish soap and kitchen sponge on her bed, or leaving a stack of her dirty dishes on her bed. Unlike Experiment 2a, we removed the two baseline conditions and focused only on figurative and literal punishments. We additionally included the opportunity for the punisher to talk to the perpetrator[6]. Our motivation behind this was two-fold: (i) it better mimicked the kinds of options we have in our daily life, enhancing the ecological validity of our experiment, (ii) it allowed for a

[6] Participants in Experiments 7a and 7b had the same three options - to talk, punish literally, or punish figuratively. The main difference was in the relationship they shared with the perpetrator. In Experiment 7a participants were described as either being friends with the perpetrator or strangers. In Experiment 7b participants were either the landlord (and the perpetrator was their tenant) or they were the tenant (and the perpetrator was their landlord).

more stringent test of the usefulness of figurative punishment (will people ever select to punish figuratively when the alternatives are to talk or punish literally?).

1.3.2.2 Results

We ran a chi-square test of goodness-of-fit and found that the proportion of choice differed across the between-subjects condition X^2 (2) = 12.03, p = .002, V_{Cramer} = 0.25, *CI:* [0.08, 0.41], n = 158 (Figure 1.5). When participants were in the condition where they had never before brought up the norm-violating behavior, 81% (*CI:* [73%, 89%]) wanted to talk to the perpetrator, 17% (*CI:* [9%, 25%]) wanted to punish her figuratively, and 3% (*CI:* [0%, 11%]) wanted to punish her literally. In contrast, when they were in the condition where they had previously spoken about the bad behavior before 60% (*CI:* [50%, 71%]) wanted to talk, 40% (*CI:* [30%, 51%]) chose to punish figuratively, and no one (0%, *CI:* [0%, 11%]) chose to punish literally.

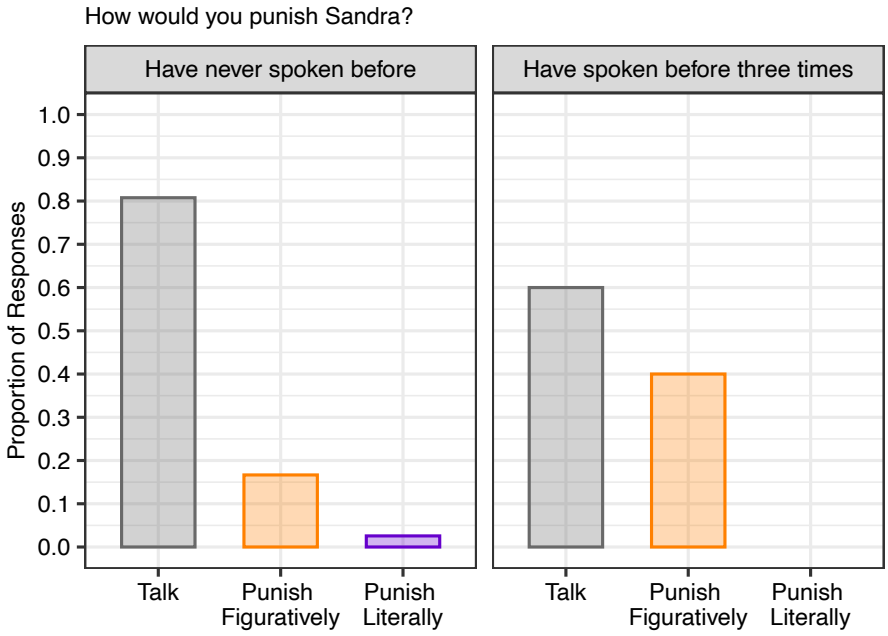

Figure 1.5: Proportions reflecting which punishment participants would choose themselves as punishers, broken down by the two experimental conditions.

1.3.3 Discussion

In Experiment 2a we presented our participants with four different kinds of punishments and asked them how they would punish the perpetrator if they were the punisher. People displayed a clear preference for costless yet communicative punishments (i.e., figurative punishments), even in the presence of costly alternatives (i.e., literal punishments). This result is difficult to reconcile with a model where the effectiveness of punishment is derived directly from its imposed incentive. However, the result is consistent with the hypothesis that human punishment is organized around communicative principles. Not only do people understand the logic of figurative punishments, they considered them to be viable and preferred alternatives to typical, cost-imposing punishments. This is most clearly evident in Experiment 2b. When participants are faced with a norm-violating agent they typically prefer to begin by talking to them about their behavior. But if this fails, their next choice is to punish figuratively, not literally. Furthermore, this pattern of choice appears consistent across two different relational contexts as evidenced from Experiments 7a and 7b (Appendix A).

An alternative interpretation for this preference, however, is that our "figurative" punishments often involve objects that are functionally appropriate for solving the underlying norm violation–for instance, a sponge can be used to do the dishes your roommate is neglecting. Possibly, then, people prefer figurative punishment not for its communicative value but for its practical value. To test for this possibility, we ran another experiment (Experiment 8, Appendix A) which was identical to Experiment 2a in every regard except that the items used for figurative punishments could not be used practically to solve the problem. For instance, instead of leaving a brand-new kitchen sponge and dish soap on her roommate's pillow, Alice leaves a toy, plastic kitchen sponge and dish soap. Changing the items to non-usable, toy items did not change our

results: across all vignettes people preferred to punish figuratively, suggesting that the preference is indeed tied to the communicative inference of the punishment.

Why might this be? We asked participants to explain their judgments, and these free responses offer some insight. Participants who chose figurative punishments did so because they judged them to communicate the transgression clearly, but also in a non-aggressive manner. They often referred to communicative goals: to let the offender know that their action was unacceptable, but in a manner that was kind, minimized the risk of future retaliation, and did not jeopardize their ongoing relationship. As one participant put it, they allowed the punisher to "kill with kindness".

1.4 Experiment 3

Experiments 1 and 2 present situations in which a figurative interpretation of costless punishment is strongly contextually favored. Experiment 3 offers a robustness check. In a situation where the context provides a salient alternative interpretation of a "costless punishment", will people still infer that it may be designed as a communicative form of sanction?

To test this, we modified the figurative punishment vignettes used in Experiments 1 and 2 so that there was a plausible explanation for the focal behavior (e.g., leaving a new sponge and dish soap on a roommate's pillow) other than the figurative punishment of a semantically related transgression. For instance, in the "dishes" vignette we told participants that the person who left these items on her roommates' pillow worked at a company that manufactured cleaning products. Additionally, rather than specifying that her roommate had not been doing her dishes, we simply referred obliquely to the fact that each roommate had various tasks to do around the house. We implemented these changes to make the intention underlying the focal behavior relatively

ambiguous, and to eliminate task demands. We then asked participants to explain why the protagonist might have performed the focal behavior in question.

1.4.1 Methods

Participants 608 people, recruited using Amazon's Mechanical Turk, completed the experiment in exchange for $0.25. After excluding those who failed to pass either of the two comprehension checks, the final sample was made up of 425 participants (30% excluded; 57% female, modal age range: 31–40 years).

Design & materials Our goal in this experiment was to determine how readily people will infer punitive communicative intent in ambiguous cases. That is, how strongly do we infer that punishment is communicative, such that we approach ambiguous acts with the possibility that they may be communicative sanctions? To this end, we presented participants with ambiguous acts that could be interpreted as figurative punishments but could also be interpreted as ordinary non-punitive acts. We asked them to explain the action in free response form in order to avoid experimenter demand.

These scenarios were modeled on those used in Experiments 1 and 2. Each scenario revolved around one agent leaving something positive (e. g., brand-new kitchen sponge) or negative (e.g., stack of dirty dishes) for another agent. Two critical changes rendered these acts ambiguous. First, we did not describe any norm violation (e.g., in the case of Sandra and Alice we mentioned that each roommate had some chores she was assigned to, but we mentioned nothing about Sandra having recently failed to do the dishes). Second, we introduced new information that provided an alternative explanation of the act (e.g., we told participants that Alice works at a company that makes sponges and dish soap). In each case, we asked our participants to explain why the item was left (see A.2, Appendix A for the list of scenarios used).

We developed a coding rubric to categorize participant responses. The rubric had 4 pre-defined categories. The first category was "Gift" explanations–i.e., those indicating a positive, gift-like inference about the item left (e.g., "I think she did that to help Sandra out with her chores. She may have seen that the old sponge was dirty or needed to be replaced and wanted to do something nice to show appreciation for Sandra.") The second category was "Disapproval" explanations—i.e., inferences that one agent was communicating disapproval of the second agent's norm violation (e.g., "I think that she left that as a subtle way of letting her know that the dishes need to be done."). The third category, "both", was used to classify those responses that explicitly mentioned both of the two previously mentioned interpretations (e.g., "Sandra might have forgotten to do the dishes when she was supposed to, or Alice was showing her appreciation for Sandra's diligence in doing the dishes regularly by buying her some new cleaning products"). Finally, the fourth category, "none", was used to classify responses that were either too vague to correctly classify (e.g., "To make sure that Sandra found them, and that she noticed they were from Alice.") or irrelevant (e.g., "Perhaps she did it to go viral for her marketing campaign."). All coding was done by two coders (one of whom was the first author) independently. Our coders agreed on 399 of the 425 responses and obtained high inter-observer reliability (unweighted $K = 0.90$, 95% *CI*, 0.86 to 0.93). Disagreements regarding the remaining 26 responses were resolved by a third coder (the senior author), who independently reviewed the coding scheme and adjudicated between the divergent codes. The full list of scenarios can be found in A.2, Appendix A. The coding scheme can be found on OSF at https://osf.io/fxdwm/.

1.4.2 Results

For costly focal behaviors (e.g., leaving a stack of dirty dishes on roommate's pillow), 81% of responses were classified under "disapproval" (*CI*: [76%, 86%]), 15% were classified under

"none" (*CI*: [10%, 20%]), 3% were classified under "gift" (*CI*: [0%, 9%]), and fewer than 1% were classified under "both" (*CI*: [0%, 6%]). See Figure 1.6.

For costless focal behaviors (e.g., leaving a brand-new kitchen sponge and dish soap on roommates' pillow), 39% of the responses were classified under "disapproval" (*CI*: [33%, 47%]), 38% were classified under "gift" category (*CI*: [31%, 45%]), 18% were classified under "both" (*CI*: [12%, 26%]), and 5% were classified under "none" (*CI*: [0%, 12%]).

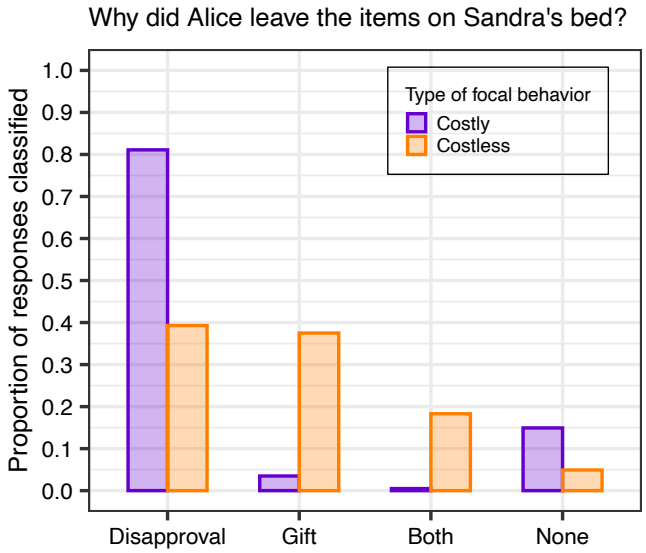

Figure 1.6: Proportion of participants free-form responses classified into one of the four predetermined categories.

1.5 Experiment 4

Our overarching goal is to understand whether punishment between humans is effective because it works as a constructed incentive or because it additionally acts as a signal of communication. Experiment 4 approaches this question from a different and complementary angle. Experiments 1, 2, and 3 ask whether benign acts can be useful punishments if they strongly convey communicative intent. In Experiment 4 we instead ask whether harmful acts can be ineffective punishments when the transparency of communicative intent is compromised by

plausible alternative motives. We generate ambiguity by considering cases of "profitable punishment", in which the act of punishment imposes a cost on the person being punished, but also generates a benefit for the person doing the punishing. Such punishments impose a direct, apparent incentive, but any intention to influence behavior through incentives is "explained away" by the profits they generate for the punisher. Specifically, each scenario in Experiment 4 described three agents: a perpetrator, a victim, and an unaffected third-party. In each case, the perpetrator committed a norm-violation against the victim and the unaffected third-party punished the perpetrator. In half of the cases, the punishment benefitted the victim. In the other half of the cases, the punishment benefitted the punisher (i.e., the unaffected third-party). In both cases, the cost imposed on the perpetrator remained exactly the same. Here is an example:

Alice, Becky, and Cassandra work together at an office. They all work in the same division and have office desks close to one another. Alice and Becky both bring home cooked lunches to work. They both store their food in the common office fridge, along with all of the other employees.

Cassandra does not bring any lunch to work. Initially, she used to buy her lunch from the office cafeteria. However, for some time now, she has been eating Becky's lunch. Everyday just before lunch time, Cassandra takes out Becky 's lunch from the common fridge and eats it without asking Becky for permission. This has been going on for quite some time and Cassandra has eaten many of Becky's lunches, without buying her any lunch in return or paying her back in any way. At first Becky didn't know who was behind her missing lunch but recently, Becky found out that the culprit is Cassandra. She has tried to speak to Cassandra about it. But Cassandra continues to eat Becky's lunch without asking her and without paying her for the food. Alice is aware of the whole situation and decides to do something about it. Alice uses Cassandra 's

credit card to buy a month's worth of groceries for Becky (or herself). She gets all of the groceries delivered to Becky (or herself) at her home and leaves the receipt on Cassandra's desk.

In both scenarios, Alice uses Cassandra's money to buy groceries; the key manipulation is whether these groceries are for Becky (Cassandra's victim) or Alice herself (a bystander). The two cases are therefore identical in the cost they impose on the perpetrator, Cassandra. They differ, however, in degree of ambiguity regarding Alice's motives. When the punishment benefits the victim (Becky), its motive is unambiguous: to dissuade Cassandra from future transgression. In contrast, when the punishment actually benefits the punisher herself, its motive is ambiguous: it may be to dissuade Cassandra from future transgression, but it may be simple self-interest.

On the simple constructed incentive model both of these punishments should be equally effective: each imposes an identical cost. In contrast, if punishment is structured to be interpreted in light of the punisher's apparent communicative intent, then "profitable punishment" by third parties should be judged ineffective, at least relative to punishment that instead compensates the victim.

1.5.1 Methods

Participants 437 people participated in the experiment. Of these, 418 completed the study and 334 (24% excluded; 55% women, modal age range was 31–40 years) passed all of the three comprehension checks. The participants were recruited through Amazon's Mechanical Turk and paid $0.35 for their participation.

Design & materials We created four new sets of scenarios (see A.3, Appendix A for the full list). Each scenario described 3 agents: a perpetrator, a victim, and an unaffected third-party. In each

case, the perpetrator transgresses against the victim and gets punished by the third-party. The punishment imposed is always costly for the perpetrator and it directly benefits either the second party (victim) or the third-party (bystander), leading to a total of 8 vignettes (scenarios (4) x beneficiary (2)). Participants were randomly assigned to read one of the vignettes, after which we asked them 2 questions, presented in the order mentioned here: "Will Cassandra get the message that she needs to stop eating Becky's lunch?" (0 = definitely not; 10 = definitely yes), and "How likely is Cassandra to stop eating Becky's lunch?" (0 = very unlikely; 10 = very likely). Each scenario included three comprehension questions and participants failing to answer any of the three correctly were excluded from all analyses.

1.5.2 Results

To analyze participant responses, we fit two linear mixed effects models (Gelman & Hill, 2006) using the lme4 package in the R statistical computing environment (Bates et al., 2015; R Core Team, 2022). For both models, there was a single dummy-coded fixed effect for beneficiary (second party (victim) = 0, third party (bystander) = 1), and random intercepts for the different scenarios.

We first analyzed participant responses to the clarity of each punishment (Figure 1.7(a)). We found a main effect of beneficiary, $b = 1.12$, $t_{329.60} = 3.09$, $p = .002$. This means that when the punisher stood to profit personally from the punishment, participants judged the punishment to be less clear and communicative to the recipient ($M = 4.96$, $SE = 0.26$) than when the exact punishment benefitted the victim of the transgression ($M = 6.15$, $SE = 0.26$).

Results for the perceived effectiveness of the punishment were similar (Figure 1.7(b)). There was a main effect of the beneficiary, $b = 0.76$, $t_{329.45} = 2.20$, $p = .03$, such that, punishments benefitting the unaffected third-party punisher were perceived to be significantly

less effective ($M = 4.67$, $SE = 0.26$) in bringing about a change in future behavior than those benefiting the affected second-party victim ($M = 5.52$, $SE = 0.25$).

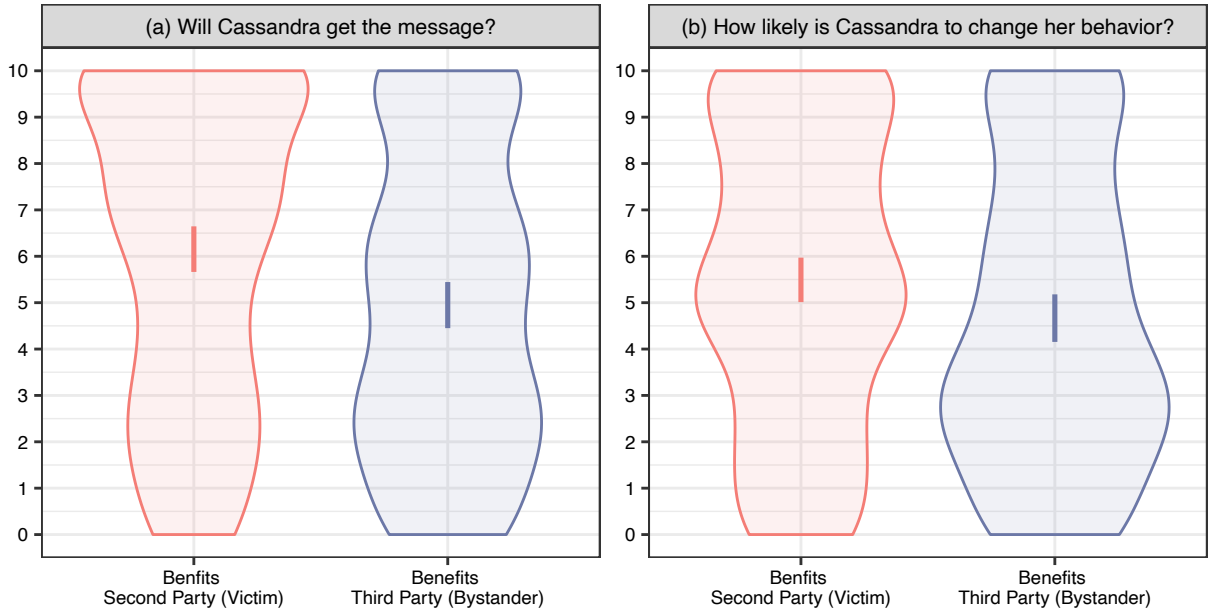

Figure 1.7: Participant response densities and 95% CIs for (a) likelihood of 'getting the message' and (b) likelihood of changing future behavior.

1.5.3 Discussion

Findings from Experiment 4 demonstrate that people expect profitable punishments to be relatively less effective at changing a perpetrator's behavior than traditional costly punishments. This is curious because the relevant "profit" and "costs" do not affect the perpetrator at all—rather, the question is whether the punishment generates profits or costs for the punisher. We can readily make sense of these data, however, by observing that the motives underlying profitable punishment are relatively less ambiguous than those underlying costly punishment. These results lend further support to the hypothesis that punishment is not best understood as a simple constructed incentive, but instead as a joint communicative and interpretive action.

1.6 General Discussion

CHAPTER 1: PUNISHMENT AS COMMUNICATION

Humans are adept at using punishment to modify each other's behavior. We ask how this works. Many current approaches model punishment as a direct imposition of cost, and model learning from punishment as a simple association between the transgression and the imposition of cost. In contrast, our research suggests that people learn from punishment by inferring the punisher's communicative intent, and that punishers construct punishments in expectation of these inferences.

We show that people expect costless, yet communicative, punishments to be as effective as typical, punitive sanctions (Experiment 1). This expectation holds across changes in perspective: people endorse the effectiveness of communicative punishments as first-party punishers (Experiment 5), as second-party recipients of the punishments (Experiment 6), and as third-party observers (Experiment 1). We find that across a range of contexts and relationships (Experiments 7a and 7b) people prefer costless punishments over more canonical cost imposing punishments, as long as the former clearly communicate the punisher's disapproval (Experiment 2a, Experiment 2b, and Experiment 8). These inferences about communicative intent occur spontaneously and even in the absence of apparent contextual cues (Experiment 3). Finally, even a punishment that imposes clear and direct costs is expected to be less effective if there is ambiguity regarding the punisher's intent (Experiment 4).

Our data thus favors a model of punishment according to which social punishment is structured not merely as an incentive, but also as a communicative act. These two dimensions of punishment are not, of course, mutually exclusive. Rather, we assume that (in addition to communicating the punisher's disapproval) an important part of what gets communicated is an implied incentive. The reason that a literal gift can act as a figurative punishment, and the reason that people expect that it will be effective, is because it is interpreted as a sanction and a threat.

Inferring and conveying communicative intent based on observable actions relies on our ability to model other people's minds (Baker et al., 2009, 2017; Wellman, 1992; Young & Saxe, 2009). In this regard it is similar to inferring and conveying the communicative intent behind language (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Grice, 1957; Kao et al., 2014; Kao & Goodman, 2015) or demonstrative action (Ho et al., 2021; Shafto et al., 2014). In order to understand instances of figurative acts as punishments, both recipients and observers must model the relationship between an actor's various possible intentions and the actions she performs, as well as the inferences that the recipient can draw based on this model.

This may involve several layers of embedded mental state inference: i.e., the punisher models the target modeling the punisher's mind. Take, for instance, the case of Alice leaving a new sponge for Sandra as "punishment". What is required for Alice to believe that Sandra will learn a lesson from this, and experience it as a punishment (as we show that participants do)? Alice must model Sandra performing an inference of communicative intent—i.e., extracting the punitive message from this act. And what is required for Sandra to perform this inference of communicative intent? Sandra must model the process by which Alice selects informative actions. Connecting these models, Alice is (choosing an action by) modeling Sandra (inferring intent by) modeling Alice's choice of action. In this way, sending and receiving the kinds of communicative signals that we demonstrate here may involve higher order theory of mind. This process is a form of pragmatic, inference-based communication (Grice, 1957; Scott-Phillips, 2014; Sperber & Wilson, 1986) that has been formalized for language (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Kao et al., 2014; Kao & Goodman, 2015) and demonstration (Ho et al., 2021; Shafto et al., 2014). Pragmatic inference grounded in natural language has been widely explored; our extension to a "pragmatics" grounded in reward and punishment indicates

how the key ideas can be extended to many other non-linguistic human behaviors. An exciting direction for future research is to explore how people make rich social inferences from relatively sparse social observations of many different kinds (see, e.g., Ho et al., 2018, 2021; Lopez-Brau & Jara-Ettinger, 2020).

More broadly, by investigating the "pragmatics of reward and punishment", the findings presented here provide a valuable starting point for drawing out further connections between moral psychology and linguistic pragmatics. For example, linguists have extensively studied how polite speech plays a key role in regulating interpersonal relationships (Brown et al., 1987) and how it requires computing and interpreting tradeoffs between communicative and social utilities (Yoon et al., 2016). Like polite speech, punishment is a mechanism by which people regulate and influence others' behaviors, and, as we argue in this paper, it can exploit inferences about communicative intent to send a message through minimally socially offensive means. Relatedly, figurative punishment provides an interesting case of indirect, non-verbal communication with parallels to "off-the-record", indirect speech such as bribing a police officer by asking "Could we deal with this ticket without going through a lot of paperwork?" (J. J. Lee & Pinker, 2010). Both of these kinds of phenomena involve communicating while delicately navigating conflicting social and moral goals, and, as such, rely on context-sensitive inferences about intentions to be successful. Future work will need to explore these rich interactions between action, communication, and morality.

Of course, there may be many instances of "literal" punishment that do not require extensive mentalizing, just as the interpretation of much literal language does not. (If you reach to grab a French fry from your partner's plate and they slap your hand away, little mentalizing is required for effective learning.) In practice, we should often be able learn from punishment by merely

tracking its "literal" incentives. Reliance on "literal" punishment may be especially common in situations where there are established norms of behavior and punishment, such as in institutional settings. Here, there are important reasons to adopt clear, consistent, and predictable sanctions for similar transgressions across individuals and across time. This may be best accomplished by direct and "proportional" sanctions for transgressions. On the other hand, mentalizing may be more essential to identifying and responding to social sanctions in less institutionalized settings—for instance, among friends and other peer groups. In these settings (as in Experiments 2a and 2b) we may often feel the imperative of sending a message without imposing an actual cost, perhaps even with plausible deniability (Lee & Pinker, 2010; Pinker et al., 2008). One reason for this could be the desire to preserve ongoing relationships. However, results from Experiments 7a and 7b fail to find evidence for increased use of figurative punishment in the context of especially close relationships. Instead, we find that people use figurative punishments with the same frequency irrespective of the relationship they share with the perpetrator (friend vs stranger; landlord vs tenant). Possibly this reflects a uniformly strong desire to preserve relationships regardless of their closeness, or possibly other motives explain the preference for figurative punishment. Resolving these possibilities is an important direction for future research. Lastly, insofar as the ability to employ and infer figurative communication is a specialized human ability, requiring higher order theory of mind, this may explain why sophisticated and subtle sanctioning systems of norms are more prevalent in humans than other species.

Why would people ever rely on actions like punishment to communicate when they could simply use language instead? In other words, why leave your roommate a sponge and dish-soap when we can much more easily just tell them, "Do your dishes!"? We consider several complementary explanations. First, action is generally more costly than speech, and thus can

serve as a more 'honest' signal of an agent's intentions and feelings (Baumeister et al., 2007; Nock, 2008). Second, communicating through sanctions may also allow punishers to circumvent potentially awkward and uncomfortable spoken interaction. Third, communicative acts may simply reinforce the contents of speech (as we see in Experiment 2b), and vice versa—two channels of communication are sometimes better than one. Finally, our ability to communicate with punishment may be evolutionarily conserved from an ancestor that did not possess natural language.

We find that figurative punishments are anticipated to be effective, and sometimes preferred. They can avoid imposing direct costs, and we know from other recent research that people often seek to enhance the communicative value and reduce the costs imposed by their punitive acts (Molnar et al., 2022). Why, then, do people also sometimes make punishments very harmful to the target (Rai et al., 2017), preferring direct, immediate incentives over figurative acts? First, as Experiment 3 suggests, in ambiguous situations actual costs support more accurate inference. Second, actual costs do not merely imply, but rather demonstrate, the punisher's willingness to incentivize their desired pattern of behavior. Third, imposing certain kinds of costs—for instance, incarceration—may incapacitate the wrongdoer, or accomplish the ancillary goal of harming them for competitive reasons (Raihani & Bshary, 2019). Finally, as noted above, institutionalized punishment requires procedural justice and impartiality, treating violations consistently across people and across time. This may favor actual costs over carefully constructed veiled threats.

Finally, how can we reconcile the apparent role of communication in punishment with evidence that people punish in one-shot, anonymous settings (Balafoutas, Grechenig, et al., 2014; Nadelhoffer et al., 2013)? We adopt two explanations that have been well explored in the

literature. First, punishment may play a communicative role that does not benefit the punisher specifically, but rather serves the victim or the community at large (Balafoutas, Grechenig, et al., 2014; Boyd et al., 2003; Boyd & Richerson, 1992; Fehr & Gächter, 2002; Gürerk et al., 2006; Henrich & Boyd, 2001) —i.e., as a form of general deterrence. Second, cognitive mechanisms supporting punishment may be adapted (by learning, or by cultural, or biological evolution) to contexts that are repeated and non-anonymous, and these mechanisms may over-extend to one-shot anonymous settings where they cannot serve those same adaptive ends (Delton et al., 2011; Jordan & Rand, 2020; Krasnow et al., 2016; Rand et al., 2014).

In sum, our work suggests re-examining a widely used, standard model of punishment. Together with other recent findings (Cushman et al., 2019; Funk et al., 2014; Ho et al., 2016, 2018; Molnar et al., 2022) it suggests that punishment is structured not only to modify the behavior of social partners by its incentive value, but also to be understood and interpreted as a form of communication. Identifying the levers of what makes punishments effective will lead to better insight into the structure and function of punishment, as well as illuminate how punishment can be leveraged to facilitate the learning of social and moral norms in humans.

2

Why punish negligence?

Abstract

Why do we often punish negligent acts?[7] Leading accounts ground punishment for negligence in two well-known, general phenomena: outcome bias and lack of due care. We isolate and verify each of these influences. Additionally, however, we find that people punish a feature unique to cases of negligence: the failure to recall relevant information at the right time. In other words, people punish mere failures of thought. Specifically, in line with the predictions of outcome bias we find that people punish negligent actors more when their negligence leads to harmful instead of harmless outcomes (Experiment 1). But they deem punishment for negligence to be successful when the actor ceases to be negligent, rather than the harm ceasing to occur (Experiment 2). Supporting the lack of due care proposal, we find that people punish negligent actors more when they fail to take reasonable antecedent precautions (Experiment 3) or when they demonstrate insufficient care towards their task (Experiment 4). Finally, supporting the

[7] This chapter is currently under review as: Sarin, A. & Cushman, F. Punishment in negligence is multifactorial: influenced by outcome, lack of due care, and the mere failure of thought.

mere failure of thought hypothesis, we find that people punish negligent actors more when they fail to bring relevant information to mind at the right time (Experiment 5), even compared to cases where the information is supplied by an exogenous reminder (Experiment 6). Taken together, our findings suggest that punishment for negligence is determined by multiple factors: bad outcome, lack of due care, and the mere failure of thought.

2.1 Introduction

Sometimes people act without sufficient thought – they forget to buy milk on their way home from work despite promising their partner they would, or they throw a dash of peanut oil in a dinner for a friend with a peanut allergy. Some of these failures end up being harmless (e.g., no milk for coffee the next morning) while others end tragically (e.g., a fatal allergic reaction). All, however, reflect instances of negligence—and, in ways ranging from small snarky comments to years in prison, many of them get punished.

When we punish peoples' negligent actions, what exactly are we punishing them for? There are several candidate possibilities. A negligent act typically has three elements (Figure 1): some insufficient antecedent preparation, a failure to recall relevant information in the moment, and a harmful consequence. In principle, any of these might be responsible for triggering our sense that punishment is due. Existing frameworks ground punishment for negligence in only the first and the third element and for good reason. The antecedent care that an agent exercises—or fails to—is a matter of intentional choice, and we know from ample research on non-negligent actions that people can be punished for their intentional choices (Cushman, 2008; Lagnado & Channon, 2008; Pillsbury, 2000; Sarin et al., 2017; Weiner, 1995). Meanwhile, there is also a wealth of evidence that people can be punished for the mere fact of causing harm, even when not intentionally chosen (Alicke, 2000; Baron & Hershey, 1988; Cushman et al., 2009; Nagel, 1979).

Both explanations, then, ground the specific case of punishment for negligence in well-established theories of punishment in general.

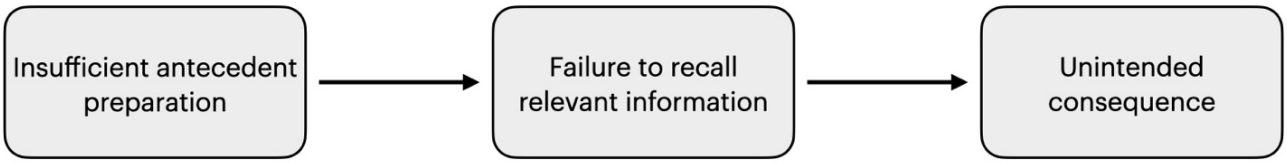

Figure 2.1: Schematic depiction of a typical negligent act.

The middle element—the very act of forgetting or failing to call relevant information to mind—has been mostly overlooked as a potential target for punishment, even though it is, arguably, negligence's most distinctive feature. If people punish the very act of forgetting, then there is a part of their punitive reaction to negligence that is special to it---a part that cannot be explained purely by relying on existing frameworks in moral psychology based on intentional acts or on harmful outcomes.

We begin by reviewing the existing frameworks in more detail, extracting testable predictions.

2.1.1 Outcome bias

Bad outcomes are like sore thumbs; they stand out from the rest and are hard to miss. And just as a sore thumb necessitates medical intervention, a bad outcome compels people to identify a culprit to hold responsible and punish. Consider a concrete case. Say Katherine is cooking dinner for friends and adds a pinch of saffron. Her friend Claire is severely allergic to saffron—something Katherine knows well, but which slips her mind. On the night of the dinner, Claire takes a bite and gets violently sick. In cases like this, the occurrence of a harmful outcome—Claire's sickness—triggers a search for the causally responsible party and assigns punishment in proportion to the severity of the harm. This holds true even if the degree of harm is caused by factors not entirely under the harm-doer's control. In the case of Claire, for instance, the severity

of her reaction will be determined in part by factors outside of Katherine's control—how big a bite Claire takes, how bad her allergy is, etc.—but, still, the amount of punishment assigned to Katherine tends to scale with the severity of Claire's reaction. This phenomenon is often called outcome bias (Baron & Hershey, 1988; Berg-Cross, 1975; Kneer & Machery, 2019; Martin & Cushman, 2016; Mazzocco et al., 2004; Walster, 1966).

Severe bad outcomes also influence our perceptions of the likelihood of the outcome occurring and the actor's desire for it, a phenomenon often called hindsight bias (Fischhoff, 2003). Accordingly, people judge bad outcomes to be more probable than benign outcomes after the fact, even when the probability of either outcome occurring, when the actor made their decision was the same (Mitchell & Kalb, 1981). They also misperceive the agent to desire the bad outcome despite being told otherwise (Fincham, 1982).

Outcome bias offers a possible explanation for why negligence is punished (Laurent et al., 2016; Margoni & Surian, 2021; Nobes et al., 2009; Nuñez et al., 2014; Shultz & Wright, 1985). In Experiment 1 we ask whether it contributes to punishment by asking whether negligent actors causing harmful outcomes should be punished more than negligent actors causing harmless outcomes. Then, in Experiment 2, we ask whether outcome bias alone is sufficient to understand the punishment of negligent acts, in which case negligent behavior that leads to harmless outcomes should not be punished.

2.1.2 Lack of due care

A second family of theories suggest that we punish negligent actors because they lack due care. What it means to 'lack care', however, is operationalized differently by different proposals within the framework.

According to one set of proposals, the lack of care in negligence refers to the actor's insufficient antecedent preparation to prevent their negligence (Amaya & Doris, 2015; Frankfurt, 1971; Murray et al., 2019; Murray & Vargas, 2020; Nobes & Martin, 2022; Sher, 2017). Thus, Katherine is deserving of punishment because she failed to take reasonable and appropriate steps to prepare for the dinner, such as texting her friends to ask about their allergies, making a mental list of allergies she knows about, checking it twice, and so on. Crucially, this family of theories holds that the capacity to exert care lies within an actor's control. When this condition is not met (due to mental limitations, distractions, or being forced to act) a person is less responsible and therefore less punishable for their negligence (Clarke, 2014b; Fincham & Jaspars, 1980; Murray et al., 2022).

A second set of proposals operationalizes the notion of 'care' as an index of an actor's values (Sher, 2017; A. Smith, 2017; Sripada, 2016; Talbert, 2017). Thus, when Katherine forgets Claire's allergy, it reveals Katherine's (insufficient) care for her. A key assumption is that when we care about and value people, information relevant to their interests (such as their allergies) will be likely to spontaneously come to mind at appropriate moments (such as when cooking for them). We will also be more likely to take intentional actions designed to prevent harm to them (e.g., texting them to ask about allergies). When these things fail to happen, it reveals the person's insufficient care, and punishment is a response to that revelation.

Although the two proposals identify and ground punishment for negligence in two different aspects of care, they are neither mutually exclusive nor entirely unrelated. If we know someone to value their friend deeply, we will predict them to take necessary precautions. Conversely, if we see someone undertaking dinner preparations for their friend with great planning and detail, we will infer that they care deeply about the friend and the dinner.

These explanations have considerable appeal because, when asked to describe negligence, ordinary people often explain it in terms of insufficient care--colloquially, carelessness (Laurent et al., 2016; Nobes et al., 2009; Nuñez et al., 2014; Shultz et al., 1986). Some experimental evidence supports this assertion. For instance, when two actors bring about a bad outcome, adults punish a careless, negligent agent significantly more than a careful agent (Margoni & Surian, 2021). Children as young as 5 distinguish between innocent and careless mistakes (Siegal & Peterson, 1998) and assign more responsibility and punishment for unintended outcomes that result from being negligent and careless than to the same unintended outcomes that stem from being negligent but careful (Nobes et al., 2009; Shultz et al., 1986). Together, this work suggests that not only do people equate negligence with carelessness (preventative and subjective), but they also find such carelessness worthy of blame and punishment.

This framework then gives us two reasons for why a negligent actor gets punished: insufficient antecedent care, and insufficient subjective care. If both play a role in how we judge negligent actors, then, the punishment given for negligence should be independently sensitive to both. Thus, actors who undertake reasonable precautions should receive milder punishment than those who fail to take those precautions. Similarly, actors who exhibit greater subjective care should get punished less than those who don't. Additionally, if these proposals are sufficient to explain punishment for negligence, then negligent actors who exhibit high subjective care or take reasonable precaution to prevent their negligence should not get punished.

2.1.3 Punishing mere failures of thought

The final element in a negligent act is its most distinctive: a person fails to bring to mind a known and relevant piece of information at the right time. Sometimes this failure leads to a bad outcome; when this happens, punishment might be explained by outcome bias. Sometimes, the

failure occurs because the negligent actor fails to take relevant precaution, or because they don't care sufficiently about the victim; when this happens, punishment might be explained by lack of due care. Although perhaps rare, there are some diagnostic cases in which, negligence occurs without leading to a bad outcome and despite the person caring deeply and taking reasonable precautions. Do people think any punishment at all is warranted in these instances? If so, this would indicate sensitivity to mere failures of thought. Moreover, if people are sensitive to failures of thought, then the mere occurrence of the relevant thought should alleviate punishment for negligence, even when the negligent actor themselves is not responsible for bringing the thought to mind–for instance, if they are reminded by somebody else.

2.1.4 Transparency & openness

Our experiments, all pre-registered[8], systematically test each of the three factors described above to assess whether they contribute to the punishment of negligence acts. To ensure that we had at least 80% power to detect the reported effects, we conducted power analyses for each experiment on pilot data collected using the reported experimental paradigm. Using mixedpower package (Kumle et al., 2023) in R Statistical Software (v4.2.1; R Core Team, 2022) (for Experiments 1 & 2) and bootstrapped power analysis (for Experiments 3, 4, 5, and 6), we carried out 1000 simulations per pilot to determine the sample size for each experiment. In all six experiments, all data were collected before any analyses were conducted. In each experiment we report on all the dependent variables collected and analyzed. Deidentified data and analysis scripts for all experiments can be found on OSF at

[8] Pre-registered documents can be found at the following links: Experiment 1 (https://aspredicted.org/3HW_1WP); Experiment 2 (https://aspredicted.org/91Z_VHN); Experiment 3 (https://aspredicted.org/QTG_QJC); Experiment 4 (https://aspredicted.org/DGB_M4S); Experiment 5 (https://aspredicted.org/PXR_R5F); Experiment 6 (https://aspredicted.org/JYD_TM5).

https://osf.io/cf7wp/?view_only=ebd456d27fe74c4d972372edc98720cb. Stimuli used for each experiment can be found in the appendices.

We recruited our participants from Amazon Mechanical Turk and paid them a small fee for their participation in the experiment. All were residents of the United States (verified by their Internet Protocol address) and gave informed consent before participating. No participant was repeated across experiment.

We collected all our data using the online labor market Amazon Mechanical Turk, which allowed us to collect data quickly and efficiently. Participants on Amazon Mechanical Turk closely match the demographic characteristics of the general U.S. population (Burnham et al., 2018). However, since our participant pool was restricted to only residents of the U.S., the approach has the limitation of restricted generalizability, and we cannot confidently generalize our findings to populations outside of the U.S.

2.1.5 Experimental logic

Experiment 1 asks whether negligent actors causing harmful outcomes are punished more than negligent actors causing harmless outcomes. Experiment 2 asks when people consider punishment given for negligence to be successful: when the harm ceases to occur, or when the actor ceases to be forgetful. Experiments 3 and 4 measure whether insufficient preparation and insufficient subjective care impact punishment judgments respectively. Experiments 5 and 6 ask whether the mere occurrence of the relevant thought mitigates punishment for negligence, even when the source of the occurrence is not the negligent actor.

2.2 Experiment 1

We begin by assessing the contribution of outcome bias to punishment for negligence. Specifically, we ask whether (1) punishment judgments for negligence display an outcome bias, and if so, whether (2) outcome bias accounts for the entirety of our punitive responses.

2.2.1 Methods

Participants An initial sample of 254 participants was recruited using Amazon's Mechanical Turk. From this, we retained a sample of 197 participants after excluding those who failed the more than 80% of our comprehension checks (n=45).

Design & materials We created a total of 36 vignettes (B.1, Appendix B) that varied an actor's mental state (intentional, negligent, accidental) and the outcome of their action (harmful, harmless) through 6 everyday contexts. Using a within-subjects balanced Latin square design, we gave each participant each of the 6 vignettes to read in random order.

Each vignette started with some background (see Figure 2.2 for a schematic overview). For instance, in one scenario, Katherine is working with a local restaurant to plan an end-of-year party for her company. As she is looking over the menu, she notices a pasta dish that has anchovies blended in it. One of the employees, Jan, is quite allergic to anchovies. In the accidental condition, Katherine *does not know* of Jan's allergy, while in the intentional and negligent conditions, she *knows* it well. A negligent Katherine, however, fails to bring the allergy to mind at the time of finalizing the menu, whereas an intentional Katherine remembers Jan's allergy. In all three conditions, Katherine approves the pasta dish with anchovies. In one half of the conditions, Jan attends the party, eats the pasta dish, and gets severely ill (harmful outcome) while in the other half of the conditions, Jan ends up giving the company party a miss (harmless outcome).

After reading each vignette, participants are asked how much punishment the actor deserves on a 10-point scale ranging from 0 (none at all) to 10 (very much). They are also asked several comprehension questions (such as: “did Katherine know that Jan is allergic to anchovies?”; “At the time of finalizing the dinner menu, did Katherine remember that Jan is allergic to anchovies?”).

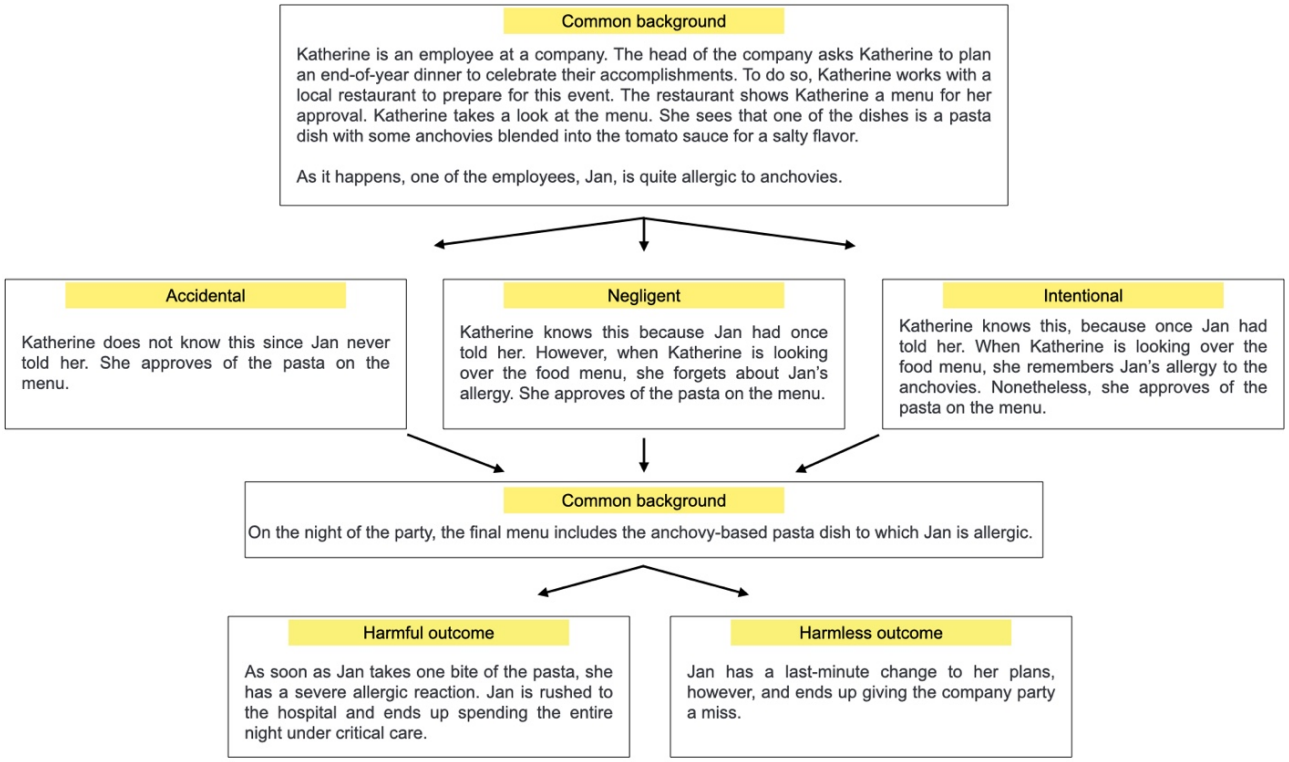

Figure 2.2: Schematic overview of one vignette for experiment 1.

2.2.2 Results

We ran a linear mixed effects model predicting participants’ punishment judgments from an interaction between mental states and outcomes adding random intercepts for participants and contexts (using lmerTest package (Kuznetsova et al., 2017) in R Statistical Software (R Core Team, 2022)). We used ‘negligent’ as the reference category for mental states and ‘harmless’ as the reference category for outcome. We found a simple effect of outcome in negligence, such

that, harmful negligent outcomes were punished significantly more than harmless negligent outcomes (b = 3.31, $t_{975.18}$ = 14.76, p < .001). We found a simple effect of mental state, such that, accidental actors (b = -1.30, $t_{975.18}$ = -5.80, p < .001) were punished significantly less than negligent actors while intentional actors were punished significantly more when no harm occurs (b = 1.21, $t_{975.18}$ = 5.40, p < .001). We also found significant interactions between mental states and outcomes, such that the difference in punishment ratings for harmful versus harmless outcomes was significantly smaller when the actor acted accidentally than when they acted negligently (b = -2.32, $t_{975.11}$ = -7.30, p < .001). Conversely the same difference was greater than that for negligent actors when actor was intentional (b = 1.18, $t_{975.11}$ = 3.72, p < .001). See Figure 2.3 for mean punishment judgments across conditions.

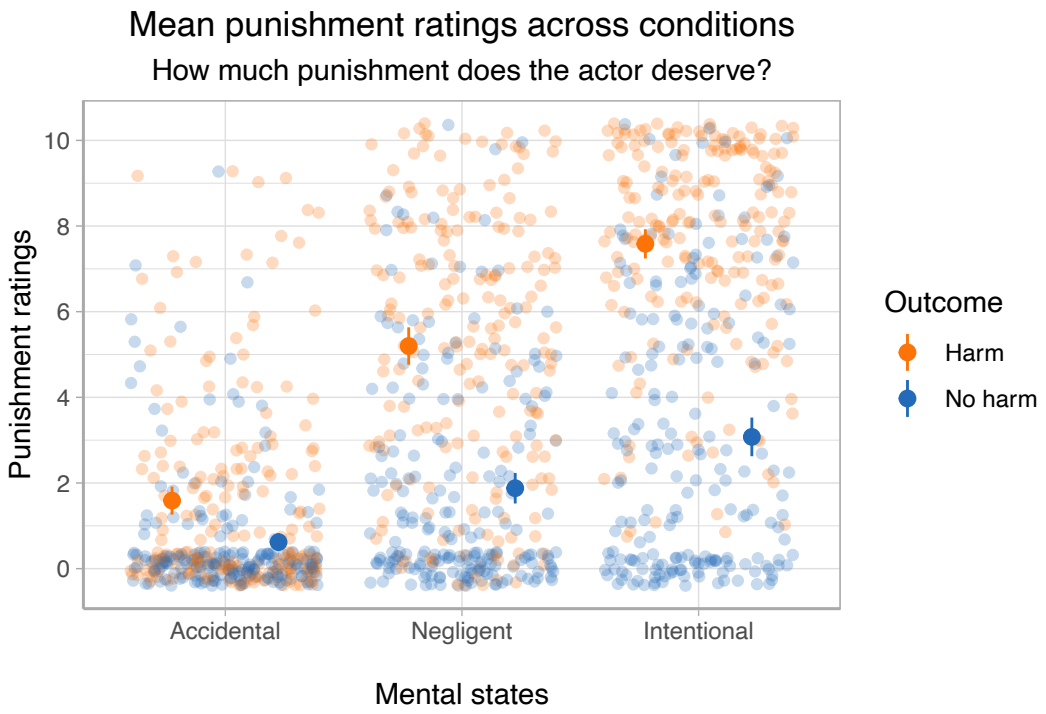

Figure 2.3: Mean punishment judgments across mental states and outcomes.

2.2.3 Discussion

In Experiment 1 we find evidence for an outcome bias: harmful outcomes garner greater punishment than harmless outcomes. The effect holds true for all mental states, including negligence, and provides us with one explanation for why we punish negligent actors: because (often) they cause bad outcomes. However, this seems to account for only part of why we punish negligence, since we also find that completely harmless acts of negligence are punished to some degree—and significantly more than similarly harmless acts done accidentally (i.e., non-negligently).

2.3 Experiment 2

In Experiment 1 we assessed the role of outcome bias in the punishment of negligence using a very standard method: By contrasting people's punitive judgments in cases that involved a harmful outcome, versus matched cases that lacked this outcome. Experiment 2 approaches the same question using a rather different methodology. All participants read about an actor who forgets a key piece of information, which leads to harm. They are asked to provide punishment ratings for the actor. They are then told that the actor faces a similar situation and either (1) again forgets the key information, but without causing harm, or (2) succeeds in remembering the key information, but accidentally causes harm. We ask whether they think the punishment was "successful". To the extent that participants rate the punishment successful in the first case, this suggests that they view the punishment as primarily aimed at harm, consistent with outcome bias. To the extent that participants instead rate the punishment as successful in the second case, this suggests that they view the punishment as primarily aimed at forgetting, or the lack of care associated with it.

2.3.1 Methods

Participants 128 participants completed the study. 17 people failed 80% of the comprehension checks were excluded from all analyses, leaving a final sample of 111. Participants were recruited through Amazon Mechanical Turk.

Design & materials The experiment had two parts. In the first part, participants read about an actor causing harm by behaving negligently, and they were asked to provide punishment judgments. This part was similar to Experiment 1, and we used four of the same contexts (using only the versions where the actor is negligent, and harm occurs). Then, in the second part, participants were told that some time had passed, and the actor was in a similar situation to before. The actor was either negligent again or not, and the outcome of their action either caused harm or didn't. Participants were then tasked to judge if prior punishment given to the actor was successful by selecting 'yes' or 'no'. Each vignette thus had one first part and four potentials second parts. Using a fully balanced within-subjects Latin-square design, each participant read about each context once – always reading the first part followed by one of the four second parts in random order (see B.2, Appendix B for the full set of vignettes).

2.3.2 Results

To see when participants think prior punishment for negligence is successful, we used a multilevel linear probability model to predict participants judgments from an interaction between mental states and outcomes, adding random intercepts for participants and contexts. We dummy-coded both variables. For mental states, being negligent was given a code of 0, while not being negligent was given 1. To test the most salient comparison between the two cases, we wanted to compare 'success' ratings between (1) cases where the actor is negligent, but no harm occurs and (2) cases where the actor is not negligent, but harm still occurs. We dummy coded our outcome variable to allow for this comparison. Accordingly, outcomes were given a code of 0 if mental

state was negligent and outcome was harmless or if mental state was not negligent and outcome was harmful, otherwise they were given a code of 1. Coding outcome in this way allowed us to get a comparison between the two salient cases as a simple effect of mental state. As predicted, we found a simple effect of mental state (b = 0.83, $t_{438.30}$ = 24.65, p < .001), indicating that participants judged prior punishment to be more successful when the actor was not negligent, even if harm occurred, compared with instances where harm ceased to occur, but the actor continued to be negligent (Figure 2.4). There was no effect of outcome (b = -0.03, $t_{439.89}$ = -0.81, p = .42) and no significant interaction between mental state and outcome (b = 0.07, $t_{439.02}$ = 1.52, p = .13).

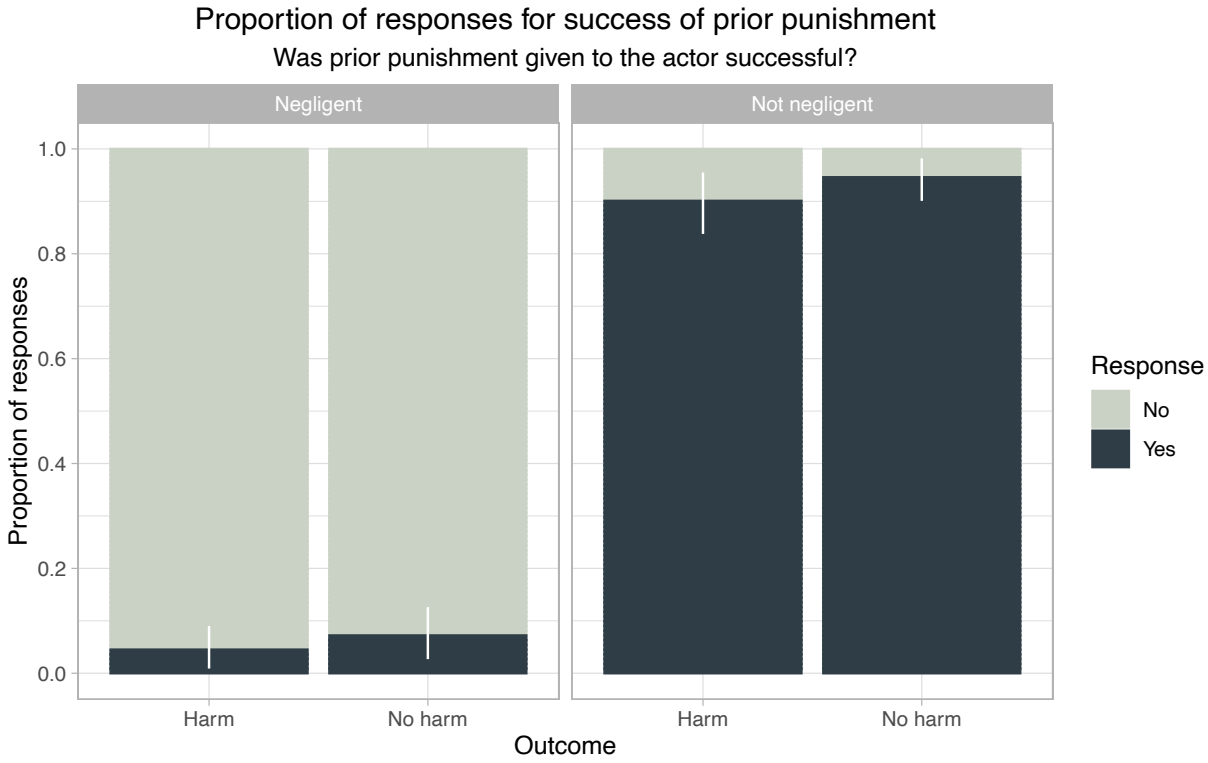

Figure 2.4: Participant responses to whether prior punishment of negligence was successful.

In addition to the multilevel linear probability model that included only random intercepts, we also ran a maximally specified model that included random slopes using brms (Bürkner,

2017). Results revealed that the Bayesian 95% credible interval did not overlap with 0 for the simple effect of mental state [0.59, 1.06]. The interval did include 0 for outcome [-0.13, 0.09], and for the interaction between mental state and outcome [-0.15, 0.29].

Finally, we asked whether participants' punishment ratings in the first part correlated with the degree to which they thought punishment was successful in the second part. To test this, we ran a pre-registered exploratory analysis, where we added participants' punishment ratings from time 1 as a predictor to our original multilevel linear probability model. Like before, we only found a simple effect for mental state (b = 0.83, $t_{437.40}$ = 24.61, p < .001). There was no effect of outcome (b = -0.03, $t_{438.96}$ = -0.83, p = .41), no outcome by mental state interaction (b = 0.07, $t_{438.28}$ = 1.54, p = .13), and no effect of participant's punishment ratings from time 1 (b = -0.001, $t_{438.44}$ = -0.38, p = .70).

2.3.3 Discussion

We asked people what counts as successful punishment for negligence: when harm ceases to occur, or when the actor ceases to be negligent (i.e., to forget important information at the necessary time). We find across various contexts that punishment of harmful negligent actions is judged to be successful only when it results in the agent ceasing to act negligently, and not when it results in the harm ceasing to occur. These results thus underscore the inference that something about negligence *per se* evokes punishment. Experiments 3 and 4 test two related explanations of what this 'something' might be.

2.4 Experiment 3

A common, often cited reason for punishing negligent actors is because they fail to act with due care. According to one set of theories, the failure signifies that a negligent actor undertook insufficient precautions ex-ante to prevent their negligence. Experiment 3 tests whether people's

punishment judgments are sensitive to this. Do people punish negligent actors more when they fail to take steps to prevent their negligence? And, if negligence occurs despite taking reasonable care, is there any punishment at all?

2.4.1 Methods

Participants From an initial sample of 741 participants, 662 completed. From these, 519 participants passed all the comprehension checks and made up the final sample. All participants were recruited using Amazon's Mechanical Turk.

Design & materials We modified the 6 contexts used in Experiment 1 to create vignettes that varied an actor's due care (sufficient vs. insufficient), leading to 12 vignettes (see B.3, Appendix B). Due care was operationalized as sufficient antecedent care. Accordingly, actors in our vignettes were described as either undertaking extensive precautions to prevent negligence, or not. Regardless of the level of care exerted, all actors (in all vignettes) end up behaving negligently, and the outcome is always harmless (thus removing any potential contribution of outcome bias to punishment). In a between-subjects design, each participant was randomly given only one vignette to read and asked how much punishment they think the negligent actor deserved on a 10-point scale ranging from 0 (none at all) to 10 (very much). Participants also answered several questions meant to check their comprehension of the context and those failing to answer any of them correctly were excluded from the analysis.

2.4.2 Results

We ran a linear mixed effects model predicting participants punishment judgments from the actor's (ex-ante) due care, adding a random intercept for contexts. We found a main effect of due care (Figure 2.5), such that, actors undertaking reasonable steps to prevent future negligence

received significantly less punishment than those who omit precaution (b = -0.57, $t_{512.45}$ = -3.04, p = 0.003).

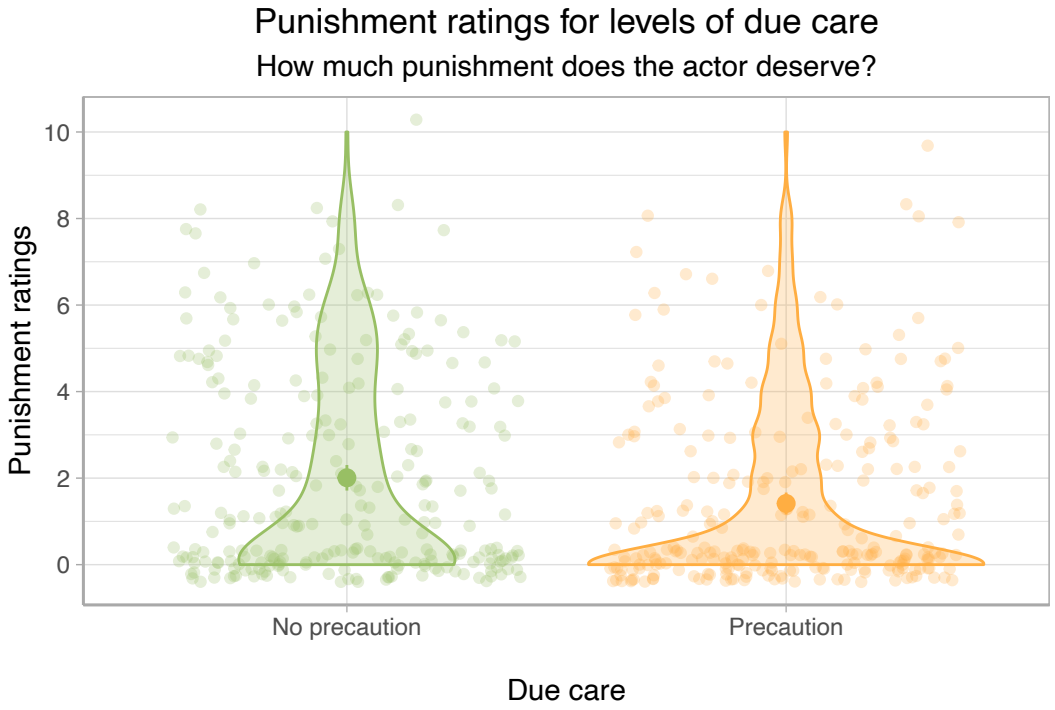

Figure 2.5: Distribution of participant responses across levels of due care (precaution). Points inside the violins represent mean punishment scores and lines represent 95% CIs of means.

2.4.3 Discussion

Experiment 3 suggests that people's punishment judgments for negligence are sensitive to how much ex-ante care and preparation a negligent actor exerts. Actors taking concrete steps to prevent future negligence, even though the negligence occurs, get punished significantly less than negligent actors who fail to take those steps. While the punishment for careful negligent actors is certainly less than careless negligent actors, it is surprising that it exists at all – after all, no bad outcome occurs, and the negligent actor deliberately takes all reasonable precautions. So why punish at all? One possibility is that when the relevant thought fails to come to mind, people might infer that the negligent actor might not subjectively care about or value the task at hand, even though they take steps to prevent the negligence. We test this in the next experiment.

2.5 Experiment 4

A second set of theories, within the due care framework, operationalizes it as a reflection of a person's underlying values. When you value someone, you remember things about them – like what they might be allergic to. Conversely then, when you negligently forget something about them, it discloses your insufficient value for them, and punishment is a response to that inference. To the extent that this proposal matters in people's judgment of negligence, variability in a negligent actor's subjective care towards another person (or task) should impact judgments of punishment. Specifically, greater subjective care should lead to lesser punishment.

2.5.1 Methods

Participants 748 people accepted the HIT posted on Amazon Mechanical Turk. Of these, 93 participants left the study incomplete and a further 165 failed to answer comprehension checks correctly. After removing these participants, we were left with a final sample of 490 participants.

Design & materials We modified the vignettes from Experiment 3 to vary an actor's subjective level of care towards the task at hand (see B.4, Appendix B). In half of the vignettes, the actor cares deeply about the task they are engaged in and wants to do a good job. In the other half of the cases, they care less and are indifferent. Like in Experiment 3, the actor always ends up being negligent, regardless of their level of care. The outcome is also always harmless. Thus, the only things that varies is the actor's degree of subjective care. In a between-subjects design, participants are randomly assigned to read a single vignette and then respond with punishment judgments (on the same 10-point Likert scale as before). They also answer comprehension checks and are excluded if they get them wrong.

2.5.2 Results

We ran a linear mixed effects model predicting participants punishment judgments from the actor's due care, adding a random intercept for contexts. We found a main effect of subjective care (Figure 2.6), such that, an actor's described as caring about the task at hand received significantly less punishment than those described as not caring (b = -1.15, $t_{484.73}$ = -5.36, p < .001).

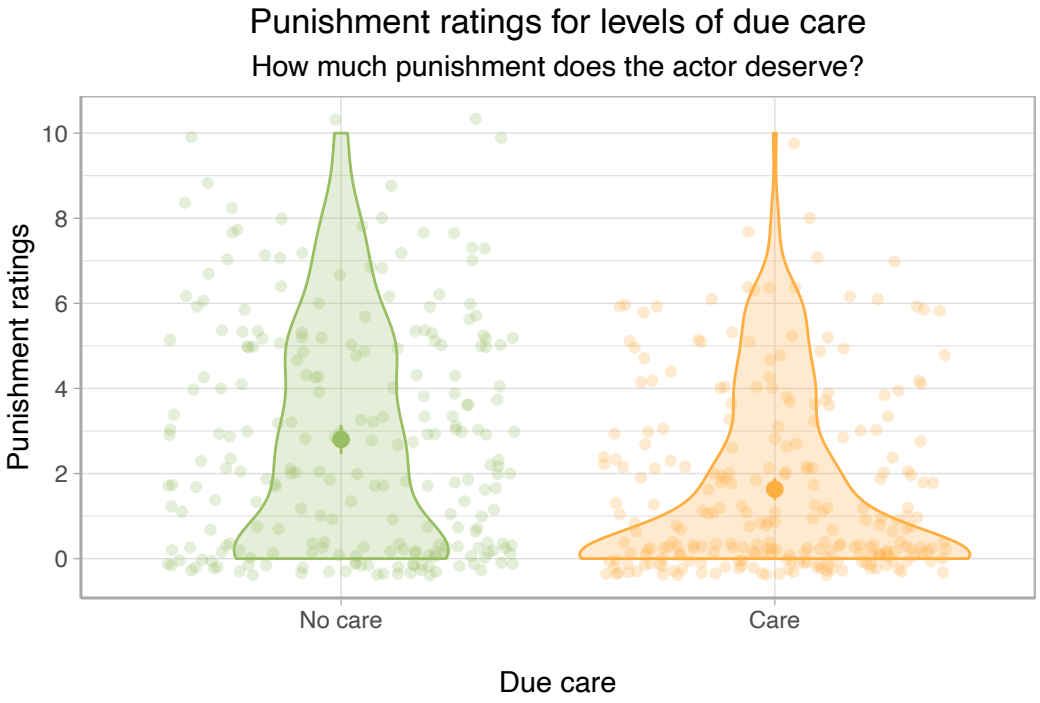

Figure 2.6: Distribution of participant responses across levels of subjective care. Points inside the violins represent mean punishment scores and lines represent 95% CIs of means.

2.5.3 Discussion

People may infer that a person's negligence stems from a deficient in their cares. In other words, if you forget about something, then, you probably didn't care about it. Consistent with this possibility, in Experiment 4 manipulating the described level of care impacted participants' punishment judgments for negligence. We find evidence in support of the subjective care hypothesis. Yet, even caring deeply about the task at hand doesn't seem to entirely absolve

punishment for negligence. This is all the more surprising given that the outcome is always harmless.

2.6 Experiment 5

This most distinctive feature of a negligent act is that it arises from a failure of thought: something that could have and should have come to mind fails to. In Experiment 5 we test whether mere failures of thought, in the presence of sufficient care and absence of harmful outcomes, elicit punishment. More specifically, we test cases of the kind where a person cares about their work, undertakes reasonable precautions, yet fails to recall an important relevant detail at the right time, thus behaving negligently. Importantly, their negligence always ends up being harmless. Thus, the only thing that varies is the mere occurrence of the relevant thought for the actor or not.

2.6.1 Methods

Participants 743 people took part in the study on Amazon Mechanical Turk. After removing those who left the study incomplete and those who failed our comprehension checks the final sample was made up of 569 participants.

Design & materials Our experiment had one independent variable of interest: the occurrence of a relevant thought at the right time. This variable had two levels: the thought either occurred, or it did not. For instance, if Katherine knows that her friend Claire is to saffron, then, in the occurrent condition, she *recalls* her friend's allergy *while making dinner for her* while in the non-occurrent condition this thought fails to come to her mind. We modified the same 6 contexts used in previous experiments to this effect, leading to a total of 12 vignettes and randomly assigned a single vignette to each participant in a between-subjects design (see B.5, Appendix B). Recall that in these vignettes no harmful outcome ever occurs (for instance, because Claire

decides to miss the dinner party) eliminating any potential role of outcome bias. In constructing these vignettes, we also specified, across all cases, that the actor exercised all reasonable precautions to reduce the influence of "due care" as a contributor to the punishment of negligent. Participants judged how much punishment the actor deserved (on a 10-point Likert scale) and answered comprehension checks.

2.6.2 Results

We predicted participants' punishment judgments from the occurrence of thought variable using a linear mixed effects model, adding random intercepts for contexts. We found a main effect of occurrence of thought (Figure 2.7) such that when the thought occurred, participants punished less (b = -1.17, $t_{566.42}$ = -9.00, p < .001).

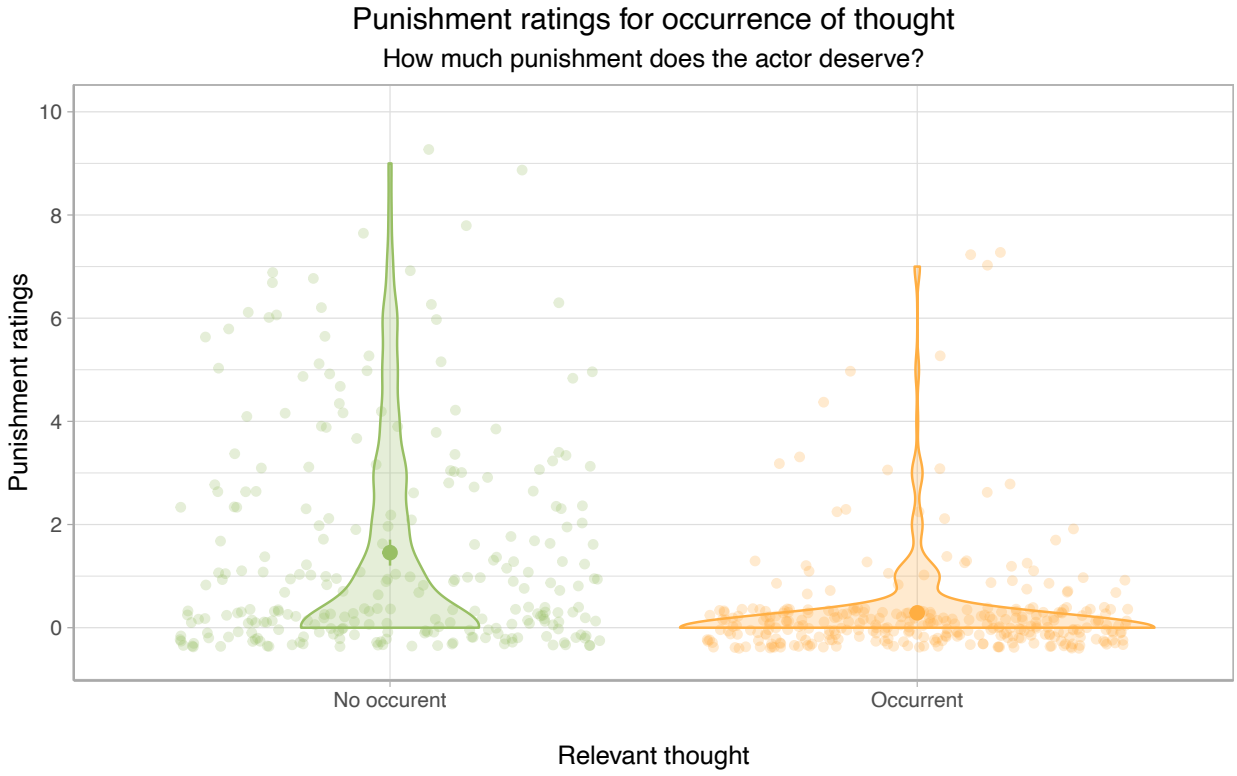

Figure 2.7: Distribution of participant responses across levels of occurrence. Points inside the violins represent mean punishment scores and lines represent 95% CIs of means.

2.6.3 Discussion

Experiment 5 asks whether a mere spontaneous failure of thought alone elicits punishment. We find that it does, even when a person exercises reasonable precautions prior to forgetting, and when no harm is caused by forgetting. In contrast, when the relevant thought does occur, punishment is at floor.

It is possible, however, that when somebody forgets information that puts another at risk, people infer that didn't care much about that other. We have already seen, in Experiment 4, that such a conclusion could drive punishment for negligence. Experiment 6 is designed to test whether this accounts for the full effect observed in Experiment 5 or whether, instead, punishment for negligence is triggered in part by direct sensitivity to the very act of forgetting something important.

2.7 Experiment 6

When a person acts negligently, it is because they have failed to call to mind something important. In Experiment 6 we sought to contrast this with the case of a person who still fails to call that thing to mind *themselves* but ends up avoiding negligence because *somebody else* reminds them. This provides us with a minimal contrast that manipulates the presence of mere negligence while holding constant all the other factors we have discussed: harm (absent in both case), sufficient antecedent preparation (present in both cases), and any potential basis to infer how much the individual cares about the forgotten information (equated across cases). After all, if we contrast two people who both fail to call information to mind, but one of whom is rescued by another person's memory, one could not draw different inferences about how much they care about the thing they forgot: Nothing about being reminded by somebody else speaks to our own degree of care.

2.7.1 Methods

Participants Of the 749 people who initially took part in the experiment, 61 left the study incomplete and 214 failed various comprehension checks. After excluding them, our final sample was made up of 474 participants.

Design & materials We took the base contexts from Experiment 5 and made one key change – in the condition where the thought occurs, the actor themselves forgets about the relevant fact but it comes to their mind because someone else reminds them. In the non-occurrent conditions, the thought never comes to mind. This lead to a total of 12 vignettes and participants are randomly assigned to read one of the vignettes in a between-subjects design (see B.6, Appendix B). After reading, they provide punishment ratings on a 10-point scale and answer a few comprehension checks.

2.7.2 Results

We use exogenous reminder to predict participant's punishment ratings using a linear mixed effects model, adding random intercepts for contexts. We find a main effect of exogenous reminder of thought (Figure 2.8), such that, when an exogenous reminder makes the relevant thought occurrent, punishment is significantly less than when the thought is not occurrent (b = -0.94, $t_{471.90}$ = -5.37, $p < .001$).

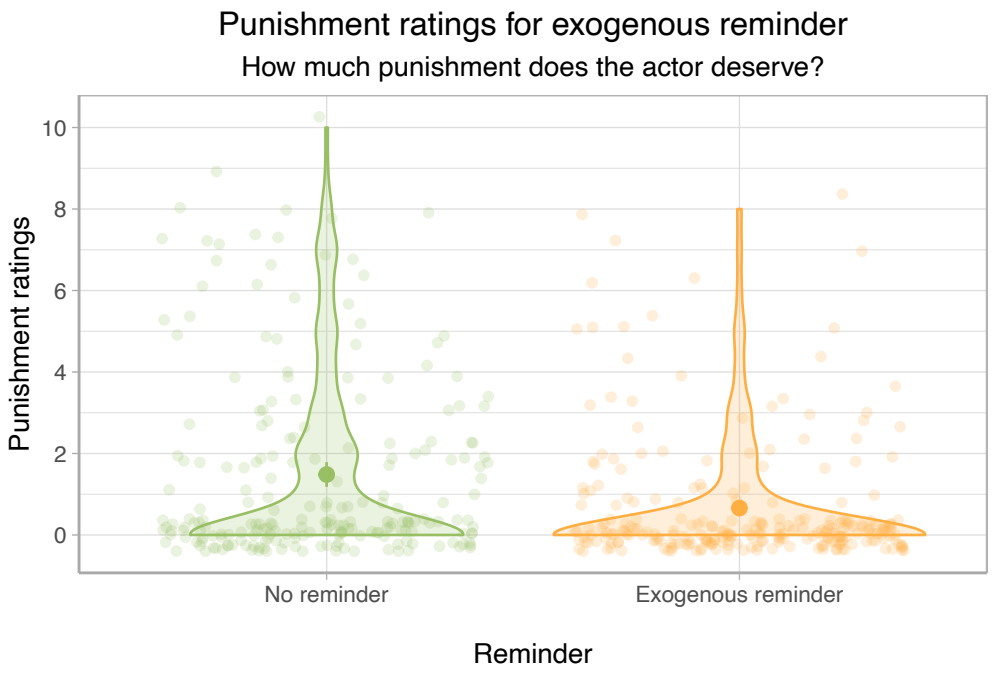

Figure 2.8: Distribution of participant responses across levels of exogenous reminder. Points inside the violins represent mean punishment scores and lines represent 95% CIs of means.

2.7.3 Discussion

In Experiment 6 we wanted to see if the punishment for mere failures of thought was due solely to an inference of lack of care. We created vignettes where the occurrence of the thought could not be attributable to the negligent actor's care. In every vignette, the actor was negligent – they failed to recall required information. However, in one condition, the relevant thought is brought to their attention through a reminder given by someone else. In the other condition, it is not. We find that people punish the actor who has the relevant thought significantly less – even though the reason this actor has the thought has nothing to do with the actor's level of care. This suggest that people may punish the mere failure of thought for its own sake.

2.8 General Discussion

People often punish others for their negligence. What, exactly, is this punishment is for? Can it be explained entirely by existing frameworks in moral psychology that focus on intentional

actions (such as antecedent care) and harmful outcomes, or is there something inherent in the negligent act itself that triggers punishment? Our results suggest that punishment for negligence is governed by multiple causes, some of which are accounted for by existing frameworks, but one of which is not.

Experiments 1 and 2 tested the role of outcome bias. Prior work shows that harmful outcomes motivate people to punish, and the resulting punishment is proportional to the degree of harm (Baron & Hershey, 1988; Kneer & Machery, 2019; Laurent et al., 2016; Shultz et al., 1981, 1986). Consistent with this, in Experiment 1, negligent actors received significantly greater punishment when their negligence resulted in harmful rather than harmless outcomes. Yet, harmless negligent acts nonetheless received significantly greater punishment than harmless acts caused by accidents, suggesting that outcome bias offers a partial explanation at best. Experiment 2 offered some insight into why: People judged prior punishment given for negligence to be successful when an actor's negligence ceased to occur instead of the harm ceasing to occur.

One explanation is that people think negligence occurs because of an actor's lack of due care (Clarke, 2014b; Murray & Vargas, 2020; Nobes & Martin, 2022; A. Smith, 2005; Sripada, 2016; Talbert, 2017). The insufficiency can be in the precautions an actor takes before the negligence occurs or it can reflect the actor's underlying values. Experiments 3 and 4 show that both kinds of care do, indeed, drive punishment. Accordingly, negligent actors receive lesser punishments when they take reasonable precautions *ex-ante* and, independently, when they care about and value the task they are engaged in and the people affected by it. Yet, even when both types of care are present and no harmful outcome occurs, negligent conduct continues to be punished.

We suggest that this residual punishment is triggered by factors quite specific to cases of negligence. Specifically, the punishment is for spontaneous failures of thought, i.e., the momentary failure to bring to mind something known and relevant to one's actions at the time. Results from Experiment 5 show that people do indeed punish mere failures of thought, even when the outcome is harmless and adequate care is present. Furthermore, the punishment is not simply because people infer lack of care from apparent failures. Rather, as results from Experiment 6 show, when a person themselves is negligent but the relevant thought is made occurrent by someone else through a reminder, punishment is reduced in comparison to when the thought is not made occurrent at all. Such an effect is difficult to explain in terms of differences in the inferences about values one would draw from negligence, since the factoring governing negligent action in these cases is exogenous to the actor themselves.

Why are people sensitive to mere failures of thought? From an adaptive standpoint, punishing such failures may serve an important function (Coricelli et al., 2005; Lieder et al., 2018; Mattar & Daw, 2018; Sarin & Cushman, 2022). Thoughts can come to mind through controlled, effortful mechanisms, but also automatically (Cushman & Morris, 2015; Dayan, 2012; Norman & Shallice, 1986). Both processes can be shaped with reinforcement, i.e., rewards and punishments (Graybiel, 2008; Lieder et al., 2018; Morris et al., 2021; O'Reilly & Frank, 2006). Getting rewarded for having the right thought at the right time would reinforce the cognitive act of bringing that thought to mind the next time the person is in a similar situation, potentially without conscious deliberation or effort. Similarly, getting punished for failing to have a thought at the right time would increase the likelihood of the omitted thought coming to mind in the future. In this way punishing momentary failures of thought can reduce a possibility

of future negligence, even in cases where negligence arises through automatic rather than controlled processes.

2.9 Conclusion

Taken together, our findings suggest that punishment for negligence is prompted by multiple sources. Echoing proposed theories in prior work we find that instances of negligence often get punished because of resulting bad outcomes or due to insufficient antecedent care taken by the negligent actor. Both proposals explain away punishment for negligence by grounding it in factors we find culpable generally (bad outcomes and careless conduct). We find that, in addition to these factors, people seem to also punish negligent actors for something unique to negligence: the very act of forgetting, of failing to bring to mind something they could have and should have. Our proposal complements the existing theories and provides a new insight into why we punish negligent actions.

3

The function of punishing negligence

Abstract

Why do we punish negligence?[9] Leading accounts explain away the punishment of negligence as a consequence of other, well-known phenomena: outcome bias, character inference, or the volitional choice not to exercise due care. Although they capture many important cases, these explanations fail to account for others. We argue that, in addition to these phenomena, there is something both fundamental and unique to the punishment of negligence itself: People hold others directly responsible for the basic fact of failing to bring to mind information that would help them to avoid important risks. In other words, we propose that at its heart negligence is a *failure of thought*. Drawing on the current literature in moral psychology, we suggest that people find it natural to punish such failures, even when they don't arise from conscious, volitional choice. Then, drawing on the literature on how thoughts come to mind, we argue that punishing a

[9] This chapter is currently under review as: Sarin, A. & Cushman, F. One thought too few: Why we punish negligence

person for forgetting will help them remember in the future. This provides new insight on the structure and function of our tendency to punish negligent actions.

3.1 Introduction

We routinely hold people responsible for their negligence, and a central goal of research in moral psychology is to understand why. Negligence can be tragic, such as when a doctor prescribes medication without checking for allergies, or when a parent forgets their child in an overheated car (Collins, 2006; Weingarten, 2010). Other times it is mundane, such as forgetting to invite a friend to a party. While instances of tragic negligence often lead to severe penalties (monetary, legal, etc.), everyday instances of negligence carry social costs of their own: a loss of face, or even a friendship.

While it seems natural for negligence to be punished, there is something profoundly strange about it as well. First, it violates certain core intuitions about the kinds of things a person can be responsible for (King, 2009; Rosen, 2004). After all, nobody "forgets" something on purpose. Negligence, in the moment that it occurs, is not a choice—any doctor, parent, or friend would have chosen to remember what they needed to. So, there is something peculiar about punishing a person for something they never really *chose*. Second, it is hard to reconcile with our best models of why punishment makes sense, i.e., its adaptive rationale. Current research emphasizes that we punish people so that they will make better choices in the future (Clutton-Brock & Parker, 1995; Fehr & Gächter, 2002; Martin & Cushman, 2016; Sarin et al., 2021). But, again, if negligence isn't really a *choice*, then how would punishing it be adaptive?

There are two classic answers to these questions. According to the first, what we punish isn't a *choice*, but an *outcome*. Merely causing harm—intentionally or not—is sufficient to trigger some amount of punishment (Baron & Hershey, 1988; Cushman, 2008; Fincham &

Jaspars, 1980; Mitchell & Kalb, 1981; Walster, 1966). Since negligent actions typically lead to harm, they are punished even though the outcome is unintended. According to the other, what we punish is not the forgetting itself, but the insufficient care taken to *prevent* that forgetting (Clarke, 2014a; Hart, 1968; Murray & Vargas, 2020; Sher, 2017). The idea is that doctors, parents, and friends are responsible for taking steps in advance—deliberately, purposely—to ensure that they won't forget their proper procedures, their children, or their guests. In other words, people are responsible for a choice they can control: To safeguard against forgetting.

Yet, both these traditions seem to miss something essential. When we hold people responsible for negligence, it seems intuitively that part of what we hold them responsible for is the very act of forgetting—a failure, in the moment, to call the right information to mind—and not just what comes after that moment (the harm) or before it (poor preparation). This is the heart of our proposal: Negligence is, above all, a *failure of thought*, in which something that could have and should have come to mind does not. First, drawing on the current literature in moral psychology, we suggest that people find it natural to punish such failures, even though they do not arise from choice. Second, drawing on a quite distinct literature on "what comes to mind"—the psychological processes that regulate what we think, and when—we argue that punishing a person for forgetting will help them remember in the future.

Specifically, current research suggests that the thoughts that come to mind are influenced by both controlled and automatic processes (Cushman & Morris, 2015; Dayan, 2012; A. Morris et al., 2019; Norman & Shallice, 1986). Crucially, both influences are responsive to reinforcement—i.e., rewards and punishments (Graybiel, 2008; Lieder et al., 2015, 2018; A. Morris et al., 2021; O'Reilly & Frank, 2006). Thus, punishing people for failures of thought can

increase the likelihood that they will automatically—without any deliberation or choice—remember what they need to in the future.

This helps us to resolve the fundamental tension presented by negligence: that we hold people responsible for failures of thought that are unintentional and beyond their control. More broadly, by grounding a theory of negligence in the mechanics of spontaneous thought and mental control, we can organize and reconcile the diverse set of factors that influence our judgments of responsibility and punishment.

3.2 Negligence

When do we hold others' responsible for negligence? In canonical cases of negligence, a person performs an action that causes harm in a way they could have foreseen but did not. For instance, they forget to buy milk on their way home from work despite promising their partner they would, or thoughtlessly add saffron to a dish while making dinner for a friend with a saffron allergy. These canonical cases have three characteristic features (Nuñez et al., 2014). First – *a lack of ill-intent*. A negligent actor does not desire the bad outcome (or the increased risk of it) that their negligence often produces[10]. Second – *awareness*. A person acting negligently acts consciously and willfully, i.e., their actions are under their control. Third – *knowledge*. A negligent actor possesses knowledge that, if appropriately considered, would make them aware of the risk of harm of their actions[11]. Accordingly, a dinner party host who adds saffron to a dish is judged to be negligent if she knows that her friend is allergic to saffron.

Many studies show that negligent behavior elicits moral condemnation and punishment, commensurate with the egregiousness of the negligence and the extremity of its consequences

[10] This separates negligent actors from intentional actors who use their knowledge and beliefs to bring about desired outcomes.
[11] This separates negligent actors from accidental actors as the latter lack pertinent knowledge.

(Laurent et al., 2016; Nobes et al., 2009; Nobes & Martin, 2022; Nuñez et al., 2014; Shultz & Wright, 1985). Our goal is to understand why.

We begin by reviewing the current literature, which offers several models of why we assign responsibility and punishment to negligence. Although these models are different in many ways, each seeks to explain how punishment for negligence can be derived from more general accounts of our "ordinary" moral judgements (i.e., moral judgments of non-negligent acts, such as purposeful harm). However, each model in its own way seems designed to avoid the simple, straightforward conclusion that accountability for negligence derives from the very act of forgetting something you ought to have remembered. Not without reason: it seems wrong to hold a person responsible for an involuntary action and forgetting is almost by necessity involuntary. It is also difficult to explain why punishing a person for something beyond their voluntary control would be adaptive.

But, while these models are important and each enjoys substantial empirical support, when taken together they seem not to offer a complete explanation. We suggest that in addition to each of the processes described in the current literature, there is something much simpler and more fundamental going on: People hold others responsible for, and punish, their acts of forgetting. The remaining challenge, then, is to explain why this would be a useful thing to do.

3.2.1 Outcome bias and moral luck

One family of theories proposes that we punish negligence because of a more pervasive "outcome bias" in moral judgment. When bad outcomes occur—harm, property damage, death, etc.—people tend to look for somebody to hold responsible, blame, and, often, to punish (Alicke, 2000; Baron & Hershey, 1988; Berg-Cross, 1975; Cushman, 2008; Cushman et al., 2009; Mazzocco et al., 2004). Moreover, the degree of punishment typically scales with the degree of

harm (Cushman, 2008; Kneer & Machery, 2019; Margoni & Surian, 2021; Nobes et al., 2009; Sarin et al., 2017; Zelazo et al., 1996). This has been described as a bias insofar as many normative theories of punishment hold that a person should not be held responsible, or punished, for events that they did not intend or that fall outside their control (Hartman, 2019; Khoury, 2018; Nagel, 1979; Nelkin, 2004; Williams & Bernard, 1981). Outcome bias is most noticeable, then, in cases like accidents, when a person causes harm unintentionally.

The same bias, however, might explain why negligent actors would be held responsible and punished. That is, they might be held responsible and punished *for the outcome* of their forgetting, rather than *for the act* of forgetting. Consistent with this possibility, negligent agents receive greater punishment when their actions lead to harmful outcomes rather than harmless outcomes (Laurent et al., 2016; Margoni & Surian, 2021; Nobes et al., 2009; Nobes & Martin, 2022; Nuñez et al., 2014; Shultz et al., 1986; Shultz & Wright, 1985).

Yet, outcome bias would seem to explain only part of the phenomenon, at best. Two arguments highlight its limitations. First, negligent actors continue to receive responsibility, blame, and punishment even when outcomes are harmless—i.e., there is no "outcome" (Amaya, 2022; Clarke, 2017; Cushman, 2008). If a customer slips and *nearly* falls on a freshly mopped floor because the cashier forgot to put up a caution sign, people might still view some degree of punishment as warranted. This clearly cannot be explained by a simple model of outcome bias. Second, intuitively, it feels that when we hold somebody responsible for negligence (even in cases that do involve bad outcomes), the thing we're holding them responsible for is not just the outcome, but the *forgetting*. Consider again the case of a person who forgets to invite their friend to their party. When the friend holds them to account for this oversight, it seems likely that they

are less motivated by the outcome (i.e., having missed the party) than the fact that they were forgotten in the first place.

3.2.2 Lack of due care and "tracing" models

A second family of theories proposes that we punish negligent actors because we can trace their negligence to antecedent choices to exercise insufficient care (Amaya & Doris, 2015; Clarke, 2014b, 2017; Frankfurt, 1971; Murray et al., 2019; Murray & Vargas, 2020; Nobes & Martin, 2022; Sher, 2009, 2017). Thus, responsibility emanates from ordinary volitional choice: The choice not to be careful enough. For instance, consider a person who knows that their friend is allergic to saffron, but forgets when preparing dinner for them, causing an allergic reaction. According to these theories, this person is responsible because they undertook preparations with insufficient thoughtfulness (for instance, they could have asked their guests about allergies, tried harder to remember any such allergies, avoided common allergens altogether, etc.).

A person's lack of due care could manifest in several ways. It can be evident through an omission, such as forgetting to invite a friend to a party (i.e., not being thoughtful enough when setting the guest list). It can be overt, such as directing one's eyes to the ingredients laid out, or covert, such as mentally running over the list of people coming to dinner and the ingredients being used. Crucially however, this family of theories avoids the difficulty of assigning responsibility to an uncontrollable act of forgetting by tracing it back to something that people *can* control: How much care they exercised.

Of course, theories like this need to put some limit on the amount of care that we can expect people to exert. These limits take several forms. We know that people are generally not held responsible in situations where they lack full mental competency or control (e.g., due to mental disturbance, usual source of distraction, having been forced to act, etc.; (Clarke, 2014a, 2014b,

2017; Fincham & Jaspars, 1980; Murray et al., 2022). In addition, philosophers have also proposed that there is an important threshold of "reasonableness" (P. Smith, 1990, 2003), which echoes a common feature of the law (Miller & Perry, 2012). It is reasonable to expect somebody to check whether the flue is open before lighting a fire, for instance; it may be unreasonable however to expect them to check whether a stranger has hidden dynamite in their chimney.

The due care proposal has several merits. First, it explains why people hold negligent actors responsible and punish them at a proximate level. It does so in familiar terms: By drawing on a theory of responsibility for things that are intentional and controllable, such as how much care you put into deciding what to do. Second, it suggests a natural explanation at the ultimate level: Holding people responsible for negligence, and punishing them, could affect their behavior by making it more likely that they will choose to exert greater care in the future. This works because exerting greater care is something over which we have clear volitional control. Finally, this view has empirical support. People often define negligence in terms of carelessness (Laurent et al., 2016; Nobes et al., 2009; Nuñez et al., 2014; Shultz et al., 1986). When two agents bring about a bad outcome, people punish a careless, (negligent) agent significantly more than a careful agent (Margoni & Surian, 2021). Children as young as 5 distinguish between innocent and careless mistakes (Siegal & Peterson, 1998) and assign more responsibility and punishment for unintended outcomes that result from being negligent and careless than to the same unintended outcomes that stem from being negligent but careful (Nobes et al., 2009; Shultz et al., 1986).

While many instances of negligence plausibly arise from a lack of due care, there are, however, other cases where people seem to assign responsibility and punishment even to agents whose degree of antecedent care was, in fact, reasonable. These suggest that the "due care" model may be incomplete.

First, consider a case in which some care was required, but the person took as much care antecedently as could reasonably have been expected—their only failing was the act of forgetting, not some antecedent preparation. For instance, imagine a person has a new puppy who needs to stay indoors. They carefully plan their return from work to keep the dog inside. But, after a stressful day of work, when they walk in the door with a bag of groceries, they end up leaving the door open for a few seconds. The puppy escapes and gets hurt. Plausibly, in this case, we hold the owner responsible not for having volitionally chosen to be insufficiently careful in advance; rather, we hold them responsible for having involuntarily forgotten*,* in the moment. Recent and ongoing research in our own laboratory confirms this: In cases like that of the dog owner, we find that people agree that the actor cared, but nevertheless hold them responsible, blameworthy, and deserving of punishment (Margoni & Surian, 2021).

Next, consider a case in which antecedent care is not reasonable to expect in the first place but, again, a person's act of forgetting leads to responsibility and punishment. For instance, imagine a person uses a space heater every day at work, and one afternoon receives an email asking them not to use electrical appliances the next day due to maintenance operations. The next day she arrives, momentarily forgets the email, and habitually turns on her space heater after putting down her bag. This results in a short circuit that cuts electricity to her whole floor. It seems natural to hold her responsible, but for what? Not her lack of *antecedent care*—it feels unreasonable to say that any such person must be expected to tape a reminder note to her door. It was reasonable for her to expect to remember. But she forgot—and it is that involuntary act of forgetting, in the moment, for which she is responsible.

There is no question that insofar as a person really does exercise less care than could have been reasonably expected *ex ante*, it is a basis for responsibility and blame. However, we suggest

that "due care" explanations may often arise, instead, from post hoc rationalization (Ditto et al., 2009; Haidt, 2001; Kunda, 1990), in order to justify a sense of responsibility that really stems from the act of forgetting. In other words, first a person involuntarily forgets, then an observer holds them responsible for it, and finally the observer goes looking for some antecedent action they *could* have taken to avoid forgetting, even if *ex ante* such an expectation of care would be unreasonable (Fischhoff, 2003; Mitchell & Kalb, 1981).

3.2.3 Character and diagnosticity

A final family of theories proposes that we hold people responsible for negligent acts because they are diagnostic of a person's underlying values (Arpaly & Schroeder, 1999; Critcher et al., 2017; Sher, 2017; A. Smith, 2005, 2017; Sripada, 2016, 2019; Talbert, 2017). A doctor who forgets to check for allergies does not value his patients sufficiently; a host who forgets a friend does not value them sufficiently, etc. This account is compatible with the idea that when we value people, we take greater care to avoid harming them, but it is also compatible with the idea that when we value people, we are simply more likely to remember them and their needs—even involuntarily—at the appropriate moment in time (A. Smith, 2005; Sripada, 2016). If you value your friends, you will take greater care to construct a guest list; if you value a *particular* friend then their name will be more likely to spontaneously come to mind in any event. When we assign responsibility and blame to a negligent act, according to these theories, what we are focused on are the values we infer from that act (or the omission).

This model is consent with evidence that many of our moral judgments are organized around an assessment of other people's character traits, which includes their values (Critcher et al., 2013; Gray et al., 2012; Malle, 2004, 2011; Pizarro & Tannenbaum, 2012; Siegel et al., 2017; Uhlmann et al., 2015).

This family of explanations suffers two shortcomings, however. First, it is overly broad: There are cases in which a person's negligence may well be diagnostic of their values, and yet we would not judge punishable. For instance, if somebody asked you to list your best friends and you forgot Alex, this might well be diagnostic of how much you value Alex. But presumably nobody would consider that act of forgetting to be *punishable*. This is likely because, as a general matter, acts that are diagnostic of a person's value do not automatically elicit the judgment that punishment is deserved. If a person orders a gold statue of themselves, tells unkind jokes about disabled household pets, or is an insufferably self-centered conversationalist, all these acts are diagnostic of a poor moral character and misaligned values. They are not, however, sufficient basis for punishment. It is challenging, therefore, to explain the punishment of negligence in terms of its diagnostic value.

Second, there are cases of the opposite type: Ones in which a person would be held responsible and punishable for negligence which is not diagnostic of their values. For instance, suppose that you have been working tirelessly on a team grant competition that would make a tremendous difference to your career, but you forget to turn it in by the deadline on the afternoon it is due. Considering all the evidence that you deeply value the grant, it would be strange for this to count as diagnostic of your values. Yet, your coworkers might hold you responsible, and your boss might sanction you. Similarly, if you take your spouse's only suit to the dry cleaners a day before they have a big meeting, they might get upset and rebuke you. Yet, their displeasure cannot be justified by any inference of your underlying values. This is so presumably because the reason you did the negligent act is presumably because you care about your spouse! Nonetheless, they might sanction you to some degree.

3.2.4 Summary: Negligence as a failure of thought

Existing models suggest three general mechanisms of moral judgment that could account for the punishment of negligent acts in particular: Outcome bias, lack of volitional due care, and character inferences. Ample evidence suggests that all three of these general mechanisms apply to negligent acts in certain cases, but each of them fails to account for the phenomenon that, we argue, is the essence of a negligent act: forgetting, or overlooking, something that ought to have come to mind.

A model that grounds negligence in a "failure of thought" makes clear contact with the folk concept of negligence as an act with no ill intent, performed with awareness, in which a person had the knowledge necessary to apprehend the riskiness of their behavior, and yet failed to do so (Laurent et al., 2016; Nuñez et al., 2014). Psychological studies of ordinary people's judgments of negligence make reference to these features as well and negligent actors are often described as culpable for failing to bring to mind something that they *could have* and *should have* (Nobes & Martin, 2022). We thus propose that this failure, to make "dispositional knowledge" (everything a person knows or could deduce) "occurrent" (i.e., consciously available to bear on decision-making) is a basic target of responsibility and punishment in humans.

As we have emphasized, however, the difficulty with this model is that it asserts we hold people responsible for something outside of their volitional control. To better understand why this might in fact be useful, we turn next to the literature on "what comes to mind"—the cognitive mechanics of remembering, consideration, and spontaneous ideation.

3.3 Consideration and negligence

In laboratory studies, the decisions we face are often quite simple: for instance, $10 today versus $11 in a week (Berns et al., 2007; Bradford et al., 2019; Kable & Glimcher, 2007; Loewenstein & Thaler, 1989). Indeed, most laboratory decision-making tasks give people a

choice between exactly two actions, each associated with one outcome. The real world, of course, is rarely so simple. We usually face many possible actions, each associated probabilistically with many possible outcomes. In other words, in the real world, there is usually too much to think about (Krajbich et al., 2010; Simon, 1955). When choosing to make financial investments, for instance, one cannot consider all possible investment vehicles, or the range of potential outcomes associated with each. Thus, cognitively efficient *consideration*—devoting attention to just the right actions and weighing just their most relevant possible outcomes—is essential to intelligent behavior (Howard & Sheth, 1969; Newell et al., 1958).

By borrowing from formal models of "consideration" in decision-making (Bear et al., 2020; Johnson & Raab, 2003; Kaiser et al., 2013; Kalis et al., 2013; Morris et al., 2019, 2021; Zhang et al., 2021), we can better understand how failures of consideration give rise to negligence. According to standard models of value-guided choice, we can understand decision-making as a situation in which an actor must choose between actions which give rise probabilistically to certain outcomes (Daw & Doya, 2006; Doya, 2008; Glimcher & Fehr, 2014; Knutson et al., 2005; Rangel et al., 2008; Rushworth et al., 2011). Each outcome is associated with some amount of reward. Normatively, the agent is expected to calculate the expected value of each action by considering all the possible outcomes of taking that action, and to choose the action with the greatest expected value. In large decision spaces, however, this is clearly not feasible. The agent thus faces two problems of efficiency in consideration—i.e., of selecting just a few relevant things to consider from among all those that are conceivable. First, they must choose a subset of possible *actions* to evaluate (Figure 3.1). Then, in order to estimate the value of each of these actions, they must choose a subset of possible *outcomes* that each of the considered actions could lead to (Morris et al., 2021).

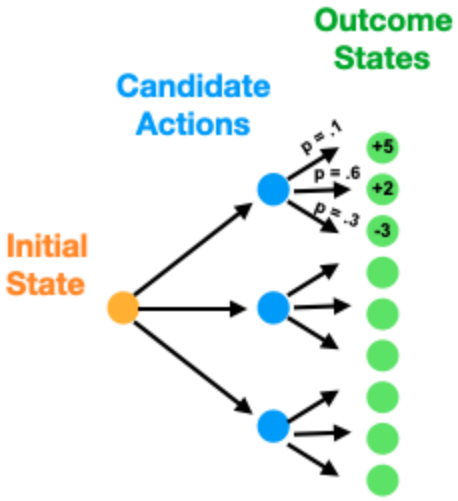

Figure 3.1: Schematic depiction of constructing a choice-set. From the space of all possible actions, people consider a few candidate actions. Each of these actions lead to several possible outcomes probabilistically. The expected utility of an action is determined by averaging over all possible outcomes.

Naturally, then, negligence can arise from a failure at either step (Figure 3.1). First, an actor can be negligent by failing to consider an action that they ought to take. For instance, a doctor might fail to consider checking the patient's allergies. Among the set of actions that they ought to have considered at a certain moment at time, an important one simply did not come to mind. Second, an actor can be negligent by failing to consider certain possible *outcomes* of the action that they select. This can lead them to miscalculate the action's expected value. In cases of negligence, this occurs when a person fails to consider the non-trivial risk of harm produced by an action they select. Examples include errors of commission such as letting your dog off leash at a park without considering that she might scare a child, or deciding to take your spouses' only suit to the dry cleaner without considering whether he will need it for work the next day.

3.3.1 Beyond actions and outcomes

The standard formal framework for decision-making—an expected value calculation over structured representations of actions and outcomes—provides a helpful start in understanding how negligence occurs. But it also has limitations. Consider again the case of failing to invite a friend to party. It feels like what has been neglected is not so much an action ("sending Alex an invitation") or an outcome ("they'll have hurt feelings"), but the *person* themselves. This is just one example of many diverse kinds of knowledge structures, other than actions and their outcomes, that can be brought to bear during decision making. We can remember the places we left our children, the ways bleach can react with other household cleaners, the list of common allergens we memorized in medical school, and so on.

When choosing actions, we often retrieve and manipulate a wide variety of representations that ultimately inform us about which actions to consider, which outcomes will occur, and how we ought to behave. These can be concrete ("Alex") or abstract ("allergens"). They can have immediate applicability to the decision at hand ("Bob is allergic to saffron"), or they can require novel, productive sequences of thought ("if bleach combines with vinegar, it could produce chlorine gas, which would kill the orchid, which would break my partner's heart"). All of these cases, however, rely on certain basic infrastructures for the regulation of thought processes, such as gating knowledge, rules, and concepts into and out of working memory (Dayan, 2012; Frank et al., 2001; Graybiel, 2008; Gruber et al., 2006; O'Reilly & Frank, 2006). When this process fails, relevant representations don't make it into working memory, leading to instances of negligence.

3.3.2 Why punish "failures of thought"?

We have argued that negligence is a failure of thought (whether an action, an outcome, or another piece of knowledge relevant to the decision process). Still, however, the question

remains why such failures would be punished, even when involuntary. One possibility is that people hold others responsible for negligent action, generally, in order to get them to improve parts of their thought processes that *are* under voluntary control, but sometimes mistakenly over-attribute negligence even to involuntary failures. In contrast, we argue that we are designed to hold people responsible for failures of thought, even those that lie beyond their control, because future thought processes are sensitive to reinforcement. To see how, we must consider more carefully how thoughts come to mind during decision making.

3.4 Automatic and controlled contributions to consideration

It takes great effort to bring certain thoughts to mind, while others appear spontaneously, or even intrude unwanted. And, indeed, there is empirical support for both modes of mental retrieval: controlled and automatic. We consider each in turn, with special attention to the role that value plays in structuring them.

3.4.1 Controlled mental retrieval

People are able to exploit their physical environment, social environment, and internal attentional mechanisms in order to call certain thoughts to mind. The role of the physical environment and social environment are easily understood: One can write a note, set an alarm, or ask your partner to remind you of something (Risko & Gilbert, 2016; Wegner et al., 1985). The ways in which internal attention can support thought retrieval are more subtle but no less important.

One widely studied mechanism is to exploit the role of structure and association between semantic representations (Kaiser et al., 2013; Kalis et al., 2013; Smaldino & Richerson, 2012). In order to remember friends, for instance, one might work through social networks ("let's think of colleagues; now schoolmates; now neighbors…"). Here, the structured relationship between

categories (which you can hold in attention) and the members of those categories (which you are attempting to retrieve) allows you to increase the likelihood of all relevant friends coming to mind. In addition to category knowledge, you can exploit episodic and contextual memory structures ("who climbed into the car this morning?"), spatial structure ("what dangerous things could Fido find in the kitchen?"), and functional structures ("what things are necessary to make a car trip successful?"), among others (Daw et al., 2011; Hauser, 2014; Hauser & Wernerfelt, 1990; Musculus et al., 2019). Of course, when directing our attention to elicit useful knowledge, the results are not entirely under our control. At each step, the semantic "prompt" that we establish (e.g., "Colleagues") can bias spontaneous thought in the direction of suitable information, but it can neither guarantee that we will direct our internal attention to all the best options and most relevant outcomes, nor can it guarantee that irrelevant or counterproductive representations will be excluded.

Still, the capacity for at least partial control over mental retrieval validates a core assumption of the "due care" model of responsibility for negligence. If a person seeking a romantic getaway arranges a picnic of campfire roasted salmon during the rainy season in Denali National Park, without even *attempting* to ask what could go wrong, they may be held accountable when, say, the raw fish sitting atop wet logs is swiped by a famished grizzly. Plausibly, what they are really being held accountable for is the failure to exert due care by querying mental associates of their plan's basic elements, which would be an intentional act of cognitive control.

Semantic structure helps us to search "widely", looking for important representations we may have overlooked. A second form of controlled mental retrieval looks "deeply" by recombining pre-existing representations into novel forms. This requires a generative process:

rules, a causal model, simulation, etc. For instance, a person might realize that it is risky to dig a hole at the base of their front step because the regular mailman is on vacation and the replacement has poor eyesight. They do not arrive at this possibility by retrieving it, precompiled—no part of them "already knew" that the replacement postman might fall in the hole. Thus, their task was not to "remember", or "retrieve". Rather, their task was to generate a new insight through combinatorial thinking—in this case, planning over a causal model ("If the postman walks up, but he can't see, then he might step right into the hole, which could really hurt him") (Botvinick & Weinstein, 2014; Daw et al., 2011; Jong & Stone, 2008; Pezzulo et al., 2018). Naturally, many tasks require both widely retrieving precompiled information and also deeply considering its implications given generative or combinatorial principles (Bornstein & Daw, 2011; Cushman & Morris, 2015; Dayan, 2012; Graybiel, 2008).

Here, again, the process of mental production will not always be entirely under our control. Although anyone in possession of the rules of addition can reliably sum two numbers, a person in possession of a causal model of their environment cannot feasibly construct all possible events arising from the digging of a hole at the base of their front step. Even when exercising care to construct many such possible events, some may be overlooked. Nevertheless, since there is at least partial control over mental production, it stands out as a second target for the "due care" model of responsibility for negligence.

3.4.2 Automaticity in mental retrieval and production: The role of value

Sometimes thoughts arise in mind spontaneously, without the application of cognitive control. When a friend asks where to get lunch, some of your favorite spots leap to mind. Or, as you arrive at your office door in the morning, it pops into your head that you never turned off the bathwater. There are also situations in which, when trying to retrieve or construct relevant

thoughts, the failure of automatic retrieval presents an irreducible obstacle. You can stand in front of an acquaintance, for instance, and make every effort possible to retrieve their name, but keep coming up with Norman (which you know is wrong) instead of Herman. Or you can stare at a construction zone and search extensively for risks without ever noticing the one that will materialize.

Several different automatic processes contribute to this spontaneous element of mental retrieval and construction. Given a semantic context, spontaneous retrieval is biased towards associates (Gigerenzer & Todd, 1999; Keller & Ho, 1988). Given a spatial or situational context, we often retrieve associated information, or simulate subsequent events, based on memory (Johnson & Raab, 2003; Kaiser et al., 2013; G. A. Klein, 1993; Mattar & Daw, 2018; Smaldino & Richerson, 2012). Information that is familiar, or routinely recalled and used, also tends to come to mind spontaneously (Bear et al., 2020; Cesana-Arlotti et al., 2012; Tversky & Kahneman, 1973).

Finally, value and reinforcement also determine what comes to mind (Bear & Knobe, 2017; Carey et al., 2019; Hitchcock & Knobe, 2009; Kaiser et al., 2013; Lieder et al., 2018; Mattar & Daw, 2018; A. Morris et al., 2021; Phillips & Cushman, 2017). This final influence, we suggest, is especially important for understanding why punishing negligent acts can be adaptive. There is evidence that value can influence which actions come to mind for consideration, which outcomes come to mind given an action under consideration, and also which other knowledge structures come to mind. But, the role that value plays in each of these cases is subtly different. We review each in turn.

3.4.3 Value-based action consideration

Candidate actions tend to come to mind specifically when they have high value—for instance, when the actions have been positively reinforced in the past (Bear et al., 2020; Kalis et al., 2013; G. Klein et al., 1995; A. Morris et al., 2019; Smaldino & Richerson, 2012). This makes sense: When trying to decide what to do, it is important to consider actions that are likely to be good (high value), but there is little value in specifically considering actions that are likely to be bad (low value). Accordingly, going on a romantic getaway would come to mind if your previous weekend excursions were romantic and relaxing but not if they were boring and lackluster (Figure 3.2). Suggestive evidence further shows that immoral options (such as stealing a new phone) are especially unlikely to come to mind (Phillips & Cushman, 2017), arguably because of their particularly low value (Phillips et al., 2019).

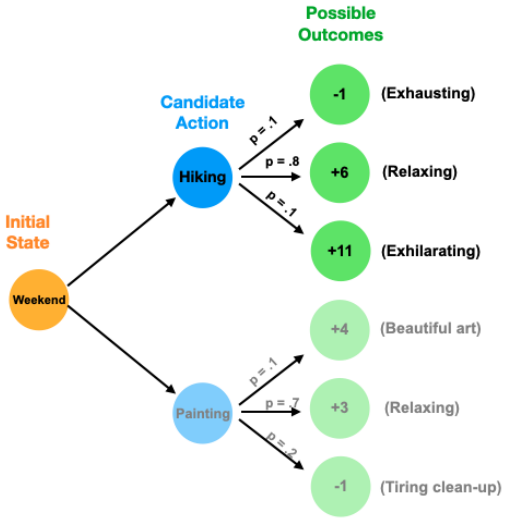

Figure 3.2: Schematic depiction of considering the expected value of performing a single candidate action by sampling a few of the possible outcomes.

3.4.4 Value-based outcome search

A single action may give rise to many different possible outcomes (Figure 3.2). Imagine that you're planning on what to do on the weekend, deliberating between two choices - to paint or go

hiking. You consider the action of going for a hike. Say that most of the hikes you've been on in the past have been relaxing and only mildly challenging. But there have been a few that have left you feeling exhilarated and on top of the world, as well as a few that have left you completely exhausted and unable to move for days. The true expected value of hiking is given by a probability-weighted average of *all* possible outcomes. These can be learned through reinforcement, but also in other ways (social learning, imagination, etc.). Crucially, however, computing the true expected value of an action is ordinarily not feasible, because of the sheer number of different outcomes that might occur (Krajbich et al., 2010; Simon, 1955). Thus, when estimating the expected value of an action, we often do so by employing a sampling approximation: We think of just a few outcomes, take a probability-weighted average of their values, and treat this as an estimate of the true expected value (Vul et al., 2014; Vul & Pashler, 2008).

Current research shows that the sampling process is biased towards sampling outcomes with high *absolute* value (Cruciani et al., 2011; Lichtenstein et al., 1978; Lieder et al., 2018; Ludvig et al., 2014; Madan et al., 2014; Mattar & Daw, 2018). In other words, outcomes that tend to be sampled (to evaluate the expected utility of an action) will be those that have been particularly beneficial and rewarding for the person, as well as those that have been particularly detrimental and distressing for the person. Utility-weighted sampling of this kind is particularly valuable because outcomes with the most extreme values dominate the overall expected value of an action (Lieder et al., 2015, 2018).

3.4.5 Other knowledge structures

When making decisions, people not only consider actions and their possible outcomes, they also can draw on a variety of other kinds of knowledge structures (who is allergic to what; the

names of friends; what happens when various detergents combine; etc.). Broadly construed, the challenge that a decision-maker faces is to select a few pieces of information from the vast quantity of knowledge they possess and to hold it in working memory, in order to suitably guide and inform choice. Here, again, current research suggests that this process of selecting, retrieving, and actively maintaining decision-relevant information is guided by value-based processes (Bear et al., 2020; Dayan, 2012; Frank et al., 2001; Gruber et al., 2006; Mattar & Daw, 2018; O'Reilly & Frank, 2006). Specifically, the "cognitive acts" involved are sensitive to reinforcement: When people experience unexpected rewards after having retrieved and used an important piece of information, this makes it more likely the same information will be retrieved again in the future (Dayan, 2012; Graybiel, 2008). For instance, when leaving the car, calling to mind the information: "My children are (or are not) in the back seat" is important. If a parent recalls this fact and narrowly avoids disaster, the reward (i.e., their extreme relief) will tend to reinforce the mental habit of checking for that information each time they leave the car in the future.

Note that the role of reinforcement in this case is slightly different that the role that value plays in the consideration of actions and outcomes (Figure 3.3). When choosing an action, *high*-value actions tend to come to mind ("How about getting ice cream?"). When considering the possible outcomes of an action, *extreme*-valued outcomes tend to come to mind (extremely high, or extremely low; "Ice cream tastes great!"; "I might spill it on my brand-new suit"). When retrieving decision-relevant information, *relevant* information tends to come to mind ("It's Labor Day, and the ice cream shops are closed"). This relevant information need not *represent* a thing of intrinsically high, low or extreme value. Rather, its relevance is a function of its value *for decision-making*.

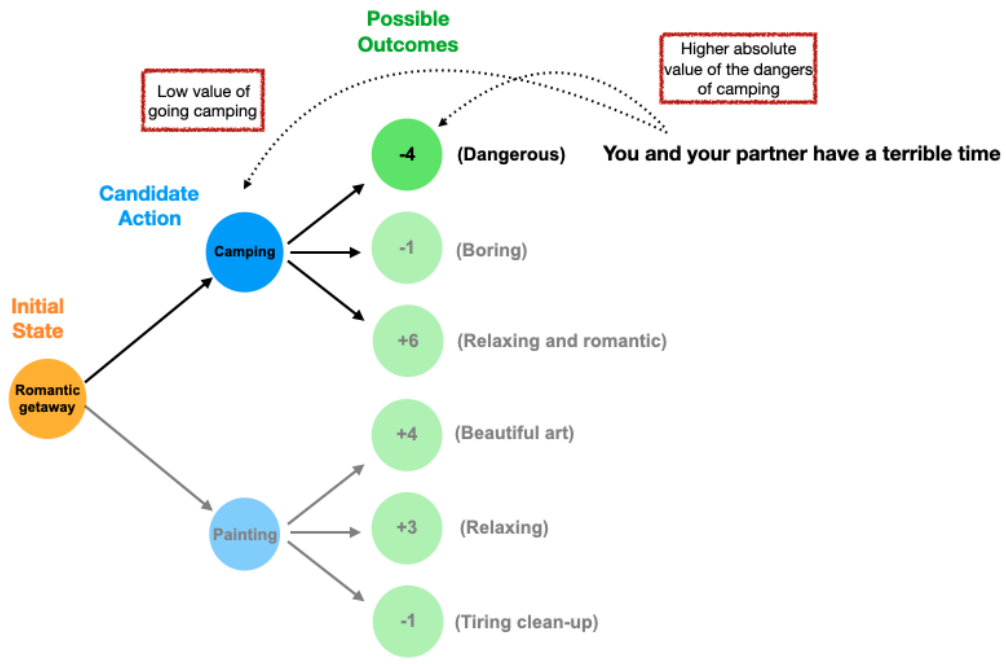

Figure 3.3: Schematic depiction of how feedback alters value of possible outcomes and candidate actions.

3.4.6 Feedback and the learning of value

Much of the literature on "what comes to mind"—actions, outcomes, and other knowledge structures—borrows formal models and methods from reinforcement learning. For this reason, it can sometimes feel as if "value" is simply what one learns from direct, personal experience. This is often the case, but there are many other ways in which humans learn value. For instance, human learning is powerfully guided by many internal and intrinsic forms of reinforcement. These include emotions like guilt, regret, or satisfaction, which may be triggered by private, offline, or quite abstract thought processes (Momennejad et al., 2018). They also include mechanisms like imagination or counterfactual reasoning (Allen et al., 2020; Gershman et al., 2017; Lohrenz et al., 2007; Lombrozo, 2020). Finally, an indispensable source of information about value is social learning (Csibra & Gergely, 2009; Herrmann et al., 2007; Ho et al., 2017).

People tell us or show us what is valuable (Ho et al., 2021; Shafto et al., 2014), we can observe their behavior and draw inferences about what is valuable (Bandura, 1962; Henrich, 2015; Olsson & Phelps, 2004; Shafto et al., 2012), and of course people also use rewards and punishments to attempt to alter each other's values (Cushman, 2013; Fehr & Gächter, 2002; Lin et al., 2012; Martin & Cushman, 2015; Sarin et al., 2021; Seymour et al., 2007; Sutton & Barto, 1998).

3.5 Synthesis

We are now in a better position to understand why punishing negligence—even negligence that arises outside of a person's intentional control—can be useful. Social punishment is a powerful form of reinforcement that will influence what comes to mind in the future. It can have this effect not only by changing people's volitional choices (consistent with the "due care" model of responsibility), but also by influencing the automatic, involuntary processes by which information comes to mind. In other words, by punishing a person for forgetting, we can influence the probability that they'll remember in the future, even if both the past and future mental acts of retrieval are automatic.

We do not suggest that ordinary people punish because they understand or explicitly represent this. Rather, at a proximate level, we suggest that people simply get frustrated at each other's negligent behavior, even when involuntary. This natural frustration, we suggest, is useful because of the role that reinforcement plays in guiding the mechanics of spontaneous thought. Also, while our emphasis has been on the interpersonal process of holding another person accountable for their negligence, a negligent person will also often experience intrinsic negative feedback in the form of regret. After all, a negligent person does not desire the harm that they end up producing. Regret is an aversive emotion that is experienced when: (i) a person makes a

choice freely and (ii) realizes (or is made aware) that an alternative choice or course of action would have produced a superior outcome than the one they choose (Bourgeois-Gironde, 2010; Chandrasekhar et al., 2008; Coricelli et al., 2005; Zeelenberg & Pieters, 2007). In other words, when a person makes a free choice only to realize that they *could have and should have* chosen differently, they experience regret. The feeling of regret in turn produces a strong negative emotional signal as well as a cognitive appraisal of one's action. This, research suggests, modifies the person's subsequent decisions and behaviors by altering the value associated with various actions (Bourgeois-Gironde, 2010; Coricelli et al., 2005; Delgado et al., 2009; Zeelenberg & Pieters, 2007). In other words, for the same reasons it can make sense for one person to hold another accountable for involuntary aspects of negligence, it can also make sense for a person to hold themselves accountable.

We can trace three distinct influences of punishing negligence on automatic processes of thought retrieval. Suppose your friend is allergic to saffron, but you forget this and feed her saffron ice cream for dessert. She has an allergic reaction, you feel badly, and your friends are shocked at the oversight. First, this negative outcome will result in a lower value estimate for the action "making saffron ice cream", which will make this *action* less likely to come to mind in the future. Second, since the value of the *outcome* (a sick friend) is extreme, this outcome will be more likely to come to mind next time you consider the same action. Finally, by representing the counterfactual "If only I had thought of Alice's allergy when cooking for her", the value of calling this information (the allergy) to mind in this context (cooking for Alice) will be reinforced, again making you more likely to avoid the same outcome again. Crucially, each of these forms of reinforcement will influence spontaneous, automatic thought processes—ones over which you have no immediate, direct volitional control.

Of course, holding people accountable for negligent actions can also influence controlled thought processes, making it more likely that people will choose to exercise "due care" in the future. For instance, when next cooking for dinner guests, you might be more likely to choose actions like writing them to check on their allergies, closely inspecting ingredient labels, etc. In theory, then, it is possible to dissociate the influence of automatic and controlled processes on information retrieval. In many real-world contexts, of course, elements of both will be blended. A person can use cognitive control to ask themselves, "what are my friends' allergies", actively maintaining and executing a search strategy ("Any gluten allergies? Any lactose intolerance?", etc.). But, during this process, the mental representations that come to mind will be influenced not only by their active, controlled search process, but also by factors over which they have no immediate, direct control, such as familiarity, frequency of use, and value.

3.6 Conclusion

Why do we punish negligence? Psychologists and philosophers have traditionally offered two answers: Outcome bias (a punitive response elicited by the *harm caused*) and lack of due care (a punitive response elicited by the antecedent *intentional choices* that made negligence possible). These factors doubtlessly contribute in many cases, and they align well with psychological models that posit causation and intention as the primary determinants of punishment (Cushman, 2008; Laurent et al., 2016; Nobes et al., 2009; Shultz et al., 1986). Another potential explanation, rooted in character-based models of moral judgment (Gray et al., 2012; Malle, 2011; A. Smith, 2017; Sripada, 2016; Uhlmann et al., 2015), is that negligence speaks to an insufficient concern for others.

These models each attempt to "explain away" negligence as an outgrowth of other, better-understood parts of our moral psychology. We have argued, however, that there is something

both fundamental and unique to negligence itself: That people simply hold others responsible for the basic fact of *forgetting* (or, more broadly, failing to call mind) things that would have made them act better. In other words, at its heart, negligence is *a failure of thought* – a failure to make relevant dispositional knowledge occurrent at the right time.

Our challenge, then, is to explain the design principles behind this mechanism of moral judgment. If we hold people directly responsible for their failures of thought, what purpose does this serve? To address this question, we draw on the literature on how thoughts come to mind. It offers a model both of how negligence occurs, and why punishing such involuntary forgetting is adaptive. Value determines which actions, outcomes, and pieces of knowledge come to mind. Specifically, actions come to mind when they have high value, outcomes when they have high absolute value, and other sorts of knowledge structures when they contribute in valuable ways to the task at hand. After an action is chosen and executed, a person receives various kinds of positive and negative feedback – environmental, social, and internal. All kinds of feedback alter value – of actions, outcomes, and other knowledge structures. Value and feedback therefore form a self-reinforcing loop: value determines what comes to mind and feedback (rewards and punishments) update value.

Our analysis sits comfortably alongside others that have already been proposed in prior work. In ordinary cases of negligent action, people seem to care about the outcome of the negligent act, the amount of antecedent deliberate care the person put in, what the negligence reveals about the person, as well as the very fact of forgetting. By drawing on theories of spontaneous thought during decision-making we can better understand the structure, function, and systematic relations among these several influences on accountability for failures of thought.

4

Punishment as instruction

Abstract

Human beings possess a natural ability for pedagogy[12]. We often utilize evaluative feedback, such as praise and criticism, or reward and punishment, as a tool to teach. Although common in daily life, there has been limited research examining how teachers teach using only evaluative feedback. Our research combines insights from Bayesian models of pedagogy and prior experimental research on evaluative feedback to address this gap. Across two experiments, we define target concepts and observe how teachers who are limited to only evaluative feedback use it to guide naive learners' choices. Our findings indicate that teachers tend to structure their feedback communicatively, in a way that minimizes uncertainty and prioritizes establishing common ground. These results offer insight into the strategies humans employ while utilizing evaluative feedback in teaching, thereby providing a more comprehensive understanding of the ease and agility with which we engage in intuitive teaching.

[12] Experiment 2 of this chapter is currently under review as: Sarin, A. & Cushman, F. Exploring teaching with evaluative feedback.

4.1 Introduction

Consider the following scenario: you have a box of delicious chocolate chip cookies sitting atop your kitchen cabinet. Your 8-year-old niece visits and is eager for a sweet treat so you decide to turn her request into a game. You inform her that a box of cookies is hidden somewhere in the house, and it is her job to find them. To do so, she must take a step in any direction and await a single-word response from you: "hot" if she is closer to the cookies or "cold" if she is further away. The game is simple enough for you to explain and easy enough for her to understand. But how, exactly, should you structure the feedback? Do you base it on her action (i.e., the direction she moved) or on her predicted movement post this action? What about her distance from the cookie – what do you say if she is physically closer to the cookies even if she can't access them, perhaps because she is standing in the adjacent room, 20-feet away? On the other hand, how does she interpret your feedback? Surely, the best way for you to structure your feedback will also depend on how you think your niece will interpret it.

Despite its beguiling simplicity, the game of hot-or-cold illustrates a sophisticated feature of our social cognition – our ability to use evaluative feedback (such as "good-or-bad"; "hot-or-cold") to shape behavior. This is a distinctive and understudied form of pedagogy. Pedagogy involves the intentional transfer of information and skills from a knowledgeable individual, the teacher, to an uninformed individual, the learner. Humans are uniquely adept at pedagogy (Csibra, 2007; Csibra & Gergely, 2006; Tomasello, 2009) and engage in it naturally and intuitively, imparting knowledge acquired through personal experience without requiring formal education (Ashley & Tomasello, 1998; Knudsen & Liszkowski, 2013).

Current research suggests that human pedagogy is often organized around principles of Bayesian inference. Specifically, a class of models implements recursive Bayesian inference and

proposes that pedagogy is a collaborative process between a teacher and a learner, where both parties are actively engaged in representing each other's mental states (Bonawitz et al., 2011; Gweon, 2021). The Bayesian inference framework formalizes rational belief updating based on observed data and can be used to predict and understand other people's behaviors, making it a useful framework for understanding pedagogy between humans. One key idea is that people have generative models of their world, which are a set of rules and representations that describe the relationship between a person's actions, their intentions, and the feedback they receive. To predict other people's behavior, we use our generative models to simulate likely actions given possible goals. Meanwhile, to understand a person's behavior, we invert our generative models to determine the most likely intentions and goals that could have produced the observed behavior. To teach, people could rely on just one of these approaches (such as, structure their feedback by relying on their generative model) or could use these approaches together (such as, structure their feedback based on their generative model, observe the learner's action post feedback and infer what the learner likely believes, and use that inference along with their generative model to select next the feedback, and so on). Models of recursive Bayesian inference represent teaching as involving both approaches, often structured hierarchically with varying levels of complexity. Such models have been used with success in understanding and predicting teaching behaviors across various goals and strategies. However, most prior work focuses on situations where teaching is initiated by the teacher, for instance by indicating relevant examples of a concept or demonstrating the use of a tool (Ho et al., 2016; Shafto et al., 2014). Less attention has been paid to pedagogical interactions where the teacher's role is restricted to providing evaluative feedback, as games like hot-or-cold illustrate.

In pedagogical interactions of this type a teacher responds with actual, verbal, or symbolic reward or punishment after a learner takes an action. Although it has not been studied much, this kind of pedagogy is common: we often use evaluative feedback to teach, and the simplicity of the feedback makes it an effective tool to teach adults, children, pets, and even robots (Fehr & Gächter, 2002; Isbell & Shelton, 2001; Owen et al., 2012). For instance, the simple phrase "good job!" can serve as an effective motivator for an adult who has completed their first coding project, a child who has finished their meal, and a pet that has successfully followed a command. The basic principle of this pedagogical interaction is also evident in games such as "hot-or-cold," where a knowledgeable player, who knows the location of the target object, guides a naive player's search by providing only "hot" or "cold" feedback.

But how does the knowledgeable player decide when to say ‘hot’ versus ‘cold’? Our study aims to understand conceptually how human teachers provide evaluative feedback. To do so, we draw upon key insights from two distinct literatures: one that demonstrates the usefulness of Bayesian inference in understanding human pedagogy, and another on the use of evaluative feedback by humans. We briefly cover these key insights in the following sections.

4.1.1 Pedagogy & Bayesian inference

Bayesian inference is a framework for probabilistic reasoning under uncertainty. It comprises of several key ideas. First, it posits that beliefs are probabilistic representations rather than point estimates—for instance, we represent the likelihood of it having rained as a distribution rather than a single setting of a binary variable (“has” or “has not”). Secondly, it provides a formal model for updating beliefs in response to new information through the application of Bayes’ theorem Eq. (4.1). This implies that we revise our posterior beliefs (e.g.,

our belief about the occurrence of rain) by incorporating our prior beliefs (i.e., the initial probability of rain) and new evidence (such as the presence of wet grass).

$$P(rain|wet\ grass) \propto P(wet\ grass|rain) * P(rain) \quad (4.1)$$

The framework has been applied successfully to model various facets of human cognition, including many instances of human pedagogy (Ho et al., 2016; Oaksford & Chater, 2009). A key insight from such work is the discovery that a Bayesian inference system implementing a recursive reasoning is particularly useful for modeling instances of human pedagogy. Models of this kind allow for iterative updating of beliefs for both the teacher and learner based on new evidence. Posterior beliefs yielded after each iteration (by combining prior beliefs with new evidence) become prior beliefs for the next iteration, allowing the framework to continually capture new information and refine its estimation of the target variable.

One notable study using this framework examined how teachers imparted a rule-based concept, specifically the boundary of a rectangle, by providing examples of points on a blank screen to a learner (Shafto et al., 2014). Rather than selecting examples in a random or arbitrary manner (selecting any point on the screen or any point within the rectangle), teachers consistently chose examples that were maximally informative, such as two positive examples at opposite ends of the rectangle, or one positive one negative example side-by-side. This pattern of example selection was in line with the predictions of a model of pedagogy that implements recursive Bayesian inference. Accordingly, teachers represent what a learner is likely to infer given various examples, and they therefore select examples that maximize a learner's belief in the correct concept (Eq. (4.2); α represents the degree to which the teacher chooses helpful examples).

$$P(d|h)_{teacher} \propto (P(h|d)_{learner})^{\alpha} \quad (4.2)$$

The learners' inferences in turn are dependent on their prior beliefs and the degree to which the examples are likely to be chosen by a helpful teacher Eq. (4.3).

$$P(h|d)_{learner} \propto P(d|h)_{teacher} * P(h) \tag{4.3}$$

This mutual dependence between teacher and learner is a key conceptual insight into the way in which a knowledgeable teacher designs instruction. Many instances of teacher-led pedagogy, such as teaching with instructions, examples, or through demonstrations operate in this manner (Ho et al., 2021; Popp & Gureckis, 2020). However, it remains to be determined whether these conceptual insights apply equally well to seemingly simpler forms of teaching, such as when teachers are restricted to providing only evaluative feedback.

4.1.2 Teaching with evaluative feedback

A simple way for us to teach others is by providing them with positive feedback when they do something good and negative feedback when they do something bad. One well-studied example is teaching with rewards and punishments. In principle, teaching with and learning from rewards and punishments could unfold in two ways: teachers and learners could use the feedback as reinforcements (meant to reinforce local action-outcome pairs), or they could engage in theory of mind and use the feedback communicatively (to communicate about the general appropriateness of an action in achieving the desired goal). Current research suggests that teachers and learners often rely on the latter strategy, provisioning evaluative feedback communicatively and interpreting it as such. For instance, when provided with feedback in the form of praise or critique, learners often tend to draw sharply different inferences about their competence based on whether the feedback came from a teacher who had prior knowledge of their ability or one who did not (Barker & Graham, 1987; Meyer, 1982, 1992; Miller & Hom, 1997). Meanwhile, when faced with deciding how to punish a transgressor, teachers often select

punishments that prioritize the message they want to communicate through their punishment, even if it means imposing less harsh penalties (Molnar et al., 2022; Sarin et al., 2021) including foregoing the punishment altogether if the message cannot be interpreted accurately due to ambiguity in the circumstances (Rai, 2022). Teachers/punishers also report being satisfied with punishment only if the learner/transgressor signals that they have understood it "as a message" and not simply experienced it as a negative incentive (Funk et al., 2014; Gollwitzer & Denzler, 2009). These findings seem consistent with the idea that when people are deciding what feedback to give, or what inferences to draw from the feedback they have received, they draw upon an explicit model of communicative intent.

Ho et al., (2018) tested this idea directly. Participants in their studies were tasked with teaching a virtual dog to walk along a path using only evaluative feedback. Across a series of experiments, dogs programmed to treat human feedback as a form of communication learnt the target path faster and more efficiently than dogs programmed to maximize human rewards, suggesting that people tend to use rewards and punishments as a form of communication (relying on our capacity for theory of mind) rather than as reinforcement (relying on our capacity for simple reward learning).

These findings offer valuable insight into how people approach teaching with evaluative feedback. However, they do so by characterizing and comparing two classes of *learner* models: reward-maximizing and communicative. As such, it leaves open the question of how teachers solve the challenge of structuring their feedback. This is the focus of the present research.

4.1.3 The present work

We take as our starting point the finding that teachers use evaluative feedback, such as rewards and punishments, communicatively rather than just as reinforcements. The question that

arises, then, is what kind of cognitive strategy they use to accomplish this. We examine this question through two experiments. In both, participants play the role of teachers, who are shown a target concept and tasked with helping a novice learner identify the concept using only evaluative feedback. Both experiments are designed to explore how teachers structure evaluative feedback from trial trial-to-trial. We consider four potential strategies teachers could employ and measure the qualitative predictions of each against data from our participants. These strategies are neither mutually exclusive, as teachers may choose to utilize them in conjunction, nor are they exhaustive, as there may exist additional ways to structure feedback. However, these strategies cover a rich conceptual space, and serve as a valuable starting point for an initial investigation into this topic.

It's worth noting that this chapter examines punishments and rewards in an educational context rather than a moralistic one. Nevertheless, it can offer a valuable template for improving our understanding of moralistic punishments. Furthermore, evaluative feedback can take several forms. For instance, non-moral evaluative feedback, such as a simple gesture like a thumbs up to acknowledge a job well done, is widely observed. Therefore, understanding the general cognitive principles that govern how we employ evaluative feedback in teaching (across both moral and non-moral contexts) can be valuable.

4.2 Experiment 1

In experiment one we designed a number line that ranged from 0-60, going up in increments of 1. Each number on the line was tied to a specific reward (1 point) except one number that had the highest reward (100 points). This number served as the target concept. Participants, who played the role of teachers, knew the identity of this number but their partners, the learners did not. The teacher’s task was to assist the learner in identifying the target concept solely through

evaluative feedback, which was given on a scale of very bad, bad, neutral, good, and very good as the learner made a series of choices.

To examine how teachers structure their feedback from trial-to-trial, we considered four cognitive strategies teachers could employ. The simplest approach for teachers is to *mirror* the reward function of the environment (which, again, is privately known to them but not the learner). This will entail giving the same feedback for all numbers on the line and the most positive feedback for the target number. This is a cognitively simple strategy, and, in the long run, it should impart an accurate understanding of the task to the learner.

Next, we consider two heuristic strategies. According to the first, the teacher constructs a *monotonic gradient* of reward defined by proximity to the target number. In other words, the closer the learner's selection is to the target number, the more reward the teacher provides. On this strategy, teacher feedback depends solely on a learner's present selection and its distance from the target and is not influenced by the learner's previous choices.

A related heuristic strategy is for the teacher to construct a *relative gradient* of reward that is defined by the learner's relative proximity to the target. According to this strategy, the teacher will give more positive feedback if the learner's current selection moves them closer to the target compared to their previous selection and negative feedback if the present selection moves them further away relatively.

The three strategies considered so far can be characterized as non-mentalizing approaches. This is so because they do not necessitate that the teacher models the learner's cognitive processes to provide feedback. Instead, a teacher may employ a simple rule or heuristic to deliver feedback without considering the potential inferences that the learner may draw from it. In

contrast, the final strategy we consider relies on the teacher's ability to model the learner's mental state.

We consider the possibility that teachers use a mental model of the learner to select the specific feedback that maximizes the learner's likelihood of subsequently choosing the target concept. This *belief-directed* strategy requires the teacher to consider the learner's mental state and the inferences they are likely to make based on the feedback received. The teacher must then select feedback that adjusts the learner's (probabilistically represented) belief towards the target concept. Giving feedback in this way would require the teacher to rely on their capacity for theory of mind and would be conceptually similar to models of human pedagogy as a form of recursive Bayesian inference.

4.2.1 Methods

Participants 500 people recruited from Amazon's Mechanical Turk took part in the experiment. After removing those who left the study incomplete, the final sample was made up of 472 participants. The majority of our participants were between the ages of 35-44 years and 35% were female-identifying.

Design & materials We constructed a number line that ranged from 0-60, going up in increments of 1 (Figure 4.1). One number on the line (number 35) was assigned a value of 100 points. All other numbers were assigned a value of 1 point each. Participants, who assumed the role of teachers, were informed that their task was to maximize the number of points earned. The most efficient means of achieving this was to select the target stimulus. However, teachers were not allowed to make selections directly; instead, they were informed that they would be paired with a "partner" (referred to henceforth as the "learner"), who could only see the number line on their

screen and not the associated rewards. The task of the teacher, therefore, was to assist the learner in earning the highest possible number of points, presumably by identifying the target number.

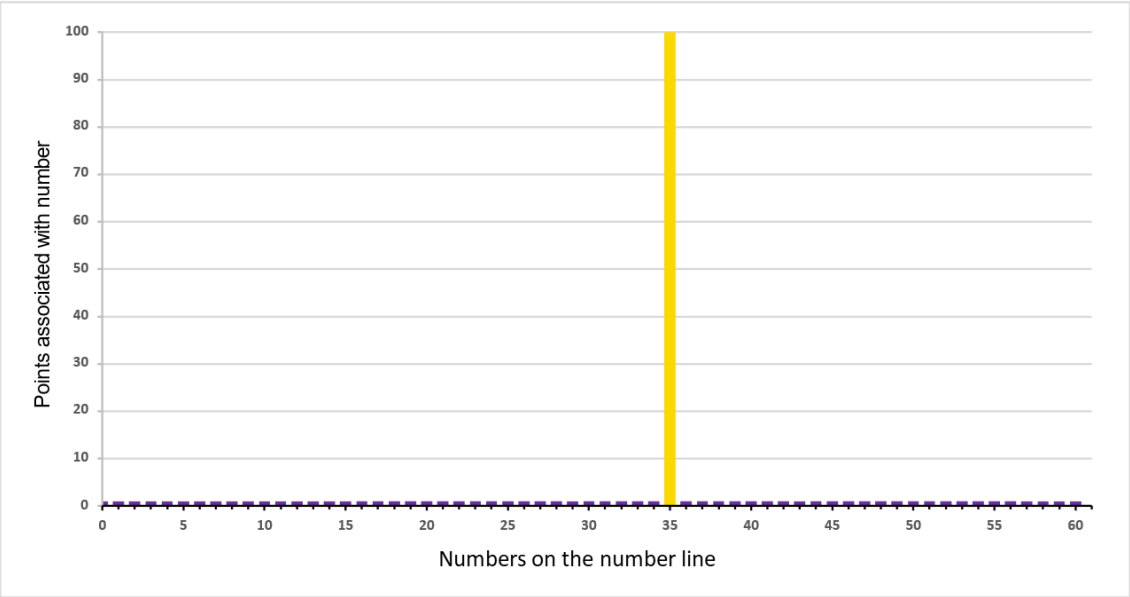

Figure 4.1: Reward distribution across the number line for Experiment 1.

On each trial, teachers were presented with the number chosen by the learner from the line and the target number and were required to provide evaluative feedback in the form of very bad, bad, neutral, good, or very good using the scale shown in Figure 4.2, for a total of 5 trials. To help quantify teacher behavior, we transformed this evaluative scale into an ordinal scale, with a range of -2 to 2.

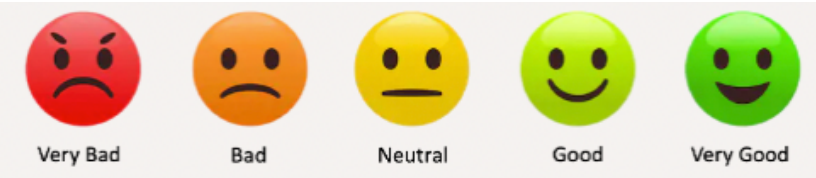

Figure 4.2: Evaluative feedback scale given to participants.

In order to compare the different teaching strategies, we programmed two distinct learner behaviors that followed a predetermined sequence of numbers. Participants interacted with only

one of the learners. Both learners chose the same numbers for the first three trials (5, 17, and 37 respectively). From trials 4 and 5, the two learners acted differently. Specifically, Learner 1 selected the number 38 followed by the number 32 while Learner 2 selected the number 32 followed by number 38. It is important to note that although the exact number the two learners selected for the last two trials differed, they were both equally spaced from the target number 35 at each trial. The two learners were programmed in a manner that would elicit unique predictions from each of the teaching strategies and facilitate comparisons within and between subjects. In the subsequent section, we delineate each strategy and detail the corresponding pattern of feedback expected to be provided by the teacher employing that particular approach.

4.2.2 Different teaching strategies: Models & predictions

Modeling Environment

Every number on the line has some reward r. The teacher's goal is to help the learner identify the number with the highest reward $max(r) = G^*$ (in this case, 35) using only evaluative feedback. Specially, on any trial t, the learner selects a number x_t and the teacher gives them feedback y_t for a total of 5 trials. Given this set up, we can define the modeling environment. If H: {0,1,2,3 ….58,59,60} is the hypothesis space for the learner, since any number on the line can be the target number, then X: $\{x_1, x_2, \dots, x_5\}$ is the set of numbers the learner chooses on any trial where $x_i \in (0, 1, 2 \dots 58, 59, 60)$. Y: $\{y_1, y_2, \dots, y_5\}$ is the feedback a teach provides in every trial where $y_i \in (-2, -1, 0, 1, 2)$. R denotes the true underlying reward distribution and is a set that contains 60 1s and 1 100. We can now define the various cognitive strategies.

Strategy: Mirroring

The number line provides rewards ranging from 1 to 100, as numbers can have a reward of either 1 or 100. In contrast, the teacher's feedback scale ranges from -2 to 2. To align the teacher's feedback with the reward function of the environment, the true scale would be converted to match the feedback scale. This conversion would result in every number having a reward of -2, except for the target number which would receive feedback of 2. Thus, a teacher using this strategy would provide the same feedback, -2, for every number on the line except the target.

The specific feedback chosen to represent the (true) reward of 1 is not critical; what matters is that the same feedback y_t is used for every number, except the target. This crucially implies that feedback will be identical for both Learner 1 and Learner 2.

$$X_{L1} = [5, 17, 37, 38, 32] \;\&\; X_{L2} = [5, 17, 37, 32, 38]$$

$$Y_{mirroring} = [y_1 = y_2 = y_3 = y_4 = y_5]$$

The resulting feedback from a teacher using this strategy to respond to our preprogrammed learners would look like a flat line, as shown in Figure 4.3.

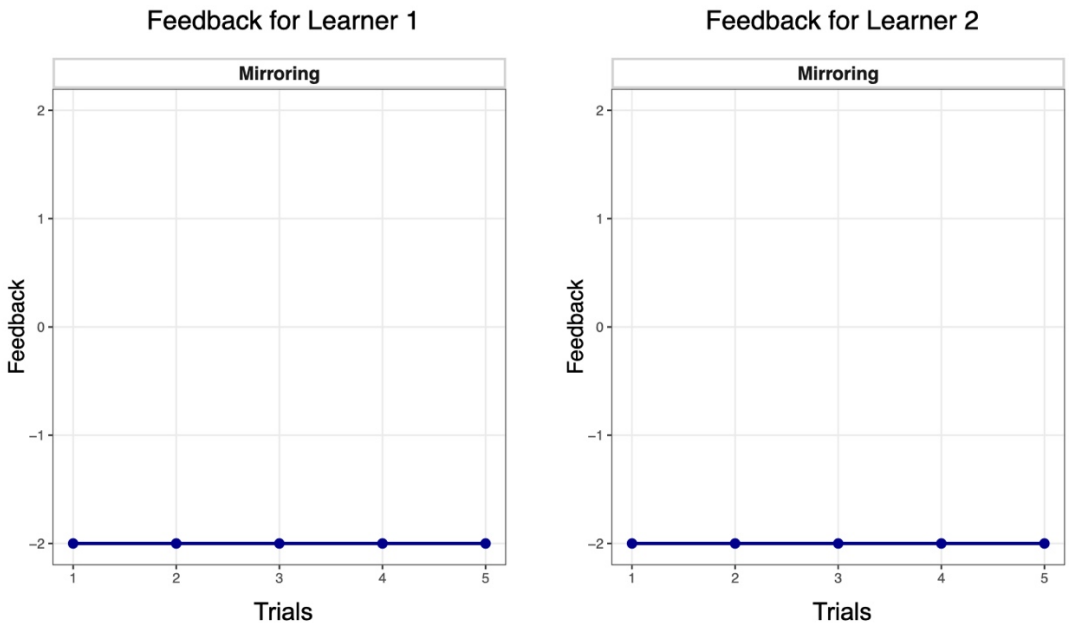

Figure 4.3: Predicted teacher feedback for Learner 1 (numbers selected: 5, 17, 37, 38, 32) and Learner 2 (numbers selected: 5, 17, 37, 32, 38) on the mirroring strategy.

Strategy: Monotonic Gradient

A teacher employing the monotonic gradient strategy would design a gradual slope or hill using evaluative feedback to assist their learner. This approach could involve associating specific evaluative feedback with sets of numbers on a line, based on their proximity to the target. For example, the teacher may divide the 36 numbers between 0 and 35 into groups of 7 and assign evaluative feedback to each group, with more negative feedback assigned to numbers farther from the target and increasingly positive feedback assigned to closer numbers. A similar gradient would be constructed between numbers 35 to 60. Under this strategy, the closer the learner's current selection x_t is to the target $G^*$the more positive the feedback they should receive on any given trial. Since selections by both Learners 1 and 2 are within a 7 number radius from the target (in trials 4 and 5), this strategy would also predict that teachers treat both learners the same.

Concretely, the feedback would be least positive for trial 1 (since x_1 is the farthest from 35 than all other numbers), relatively more positive for trial 2, and the most positive for trials 3, 4, and 5. Importantly, feedback would be identical for the last 3 trials (see Figure 4.4). So, for:

$$X_{L1} = [5, 17, 37, 38, 32] \;\&\; X_{L2} = [5, 17, 37, 32, 38]$$

$$Y_{monotonicgradient} = [y_1 < y_2 < y_3 = y_4 = y_5]$$

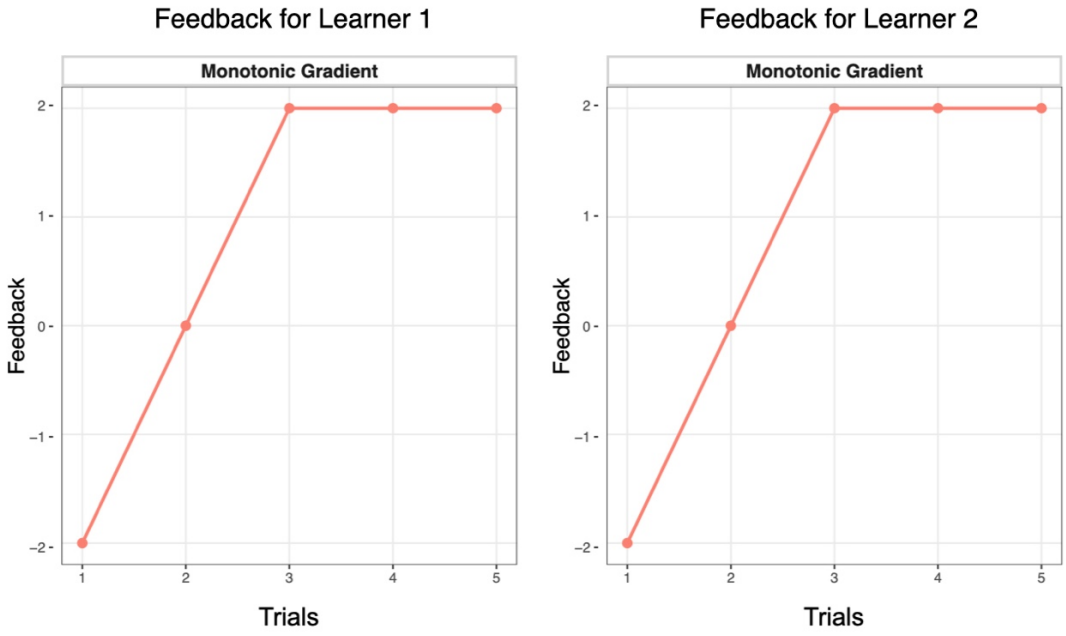

Figure 4.4: Predicted teacher feedback for Learner 1 and Learner 2 on the monotonic gradient strategy

Strategy: Relative Gradient

A teacher structuring their feedback using a relative gradient strategy focuses not only on the learner's current proximity to the target, but also their relative proximity based on past selections. To clarify, the teacher considers whether the learner's current choice x_t moves them closer to or farther from the target number G^*in comparison to their previous selection, x_{t-1}. We suggest the following equation as a means for the teacher to determine how to provide feedback:

$$Y_t = \frac{|G^* - x_{t-1}| - |G^* - x_t|}{|G^* - x_{t-1}|} \times 2 \tag{4.4}$$

A teacher following this strategy to give feedback would give increasingly positive feedback for trials 1 to 3, since both learners get relatively closer to the target on each trial. Following Eq. (4.4), the feedback would be 0, 1, and 2 respectively. On trial 4 feedback would become negative since both learners move relatively further away from the target number. However, the relative movement of both learners is the same since they move from a number (37) that is two numbers away from the target to one that is three away (38 for Learner 1, 32 for Learner 2). Based on Eq. (4.4) this would result in feedback of -1. Finally, on trial 5, both learner's relative movement is the same and their new distance from the target is the same as their old distance from the target ($|G^* - x_{t-1}| - |G^* - x_t| = 0$), prompting feedback of 0 (predictions depicted in Figure 4.5). So, for:

$$X_{L1} = [5, 17, 37, 38, 32] \;\&\; X_{L2} = [5, 17, 37, 32, 38]$$

$$Y_{relativegradient} = [0, 1, 2, -1, 0]$$

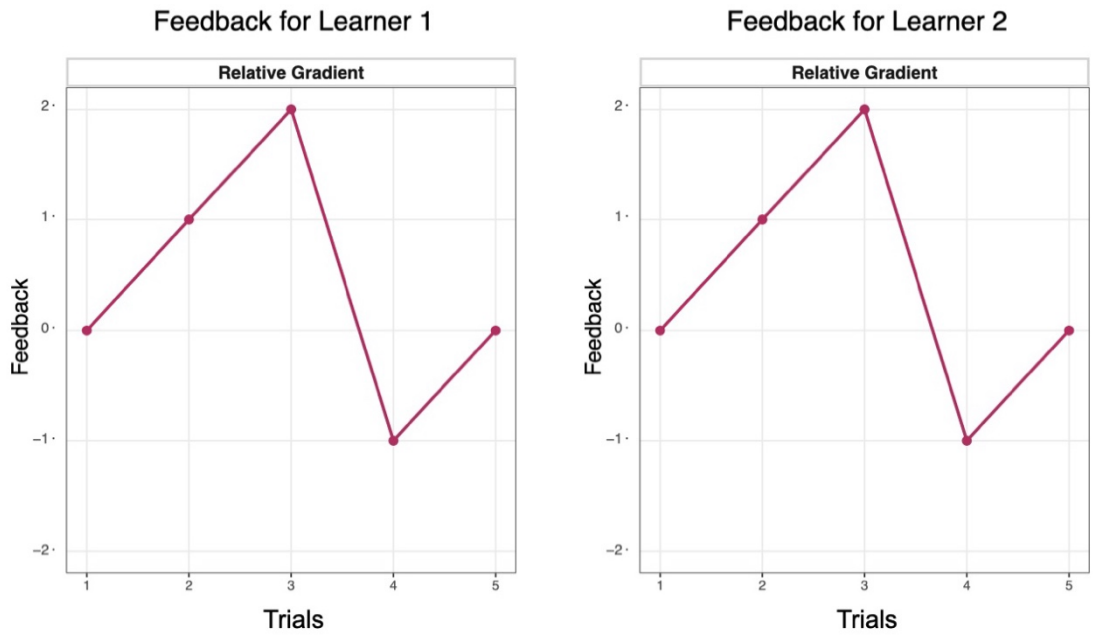

Figure 4.5: Predicted teacher feedback for Learner 1 and Learner 2 on the relative gradient strategy

Strategy: Belief directed

A teacher using a belief-directed strategy tailors their feedback to help the learner identify the target number by modeling the learner's thought process. By considering the number the learner selects on a given trial and their selection history, the teacher anticipates the learner's likely inferences regarding the identity of the target number, for different feedback types and then selects the feedback that is most likely to increase the learner's belief about the target number (Figure 4.6). This approach can be formalized as recursive Bayesian inference (Ho et al., 2018; Shafto et al., 2014).

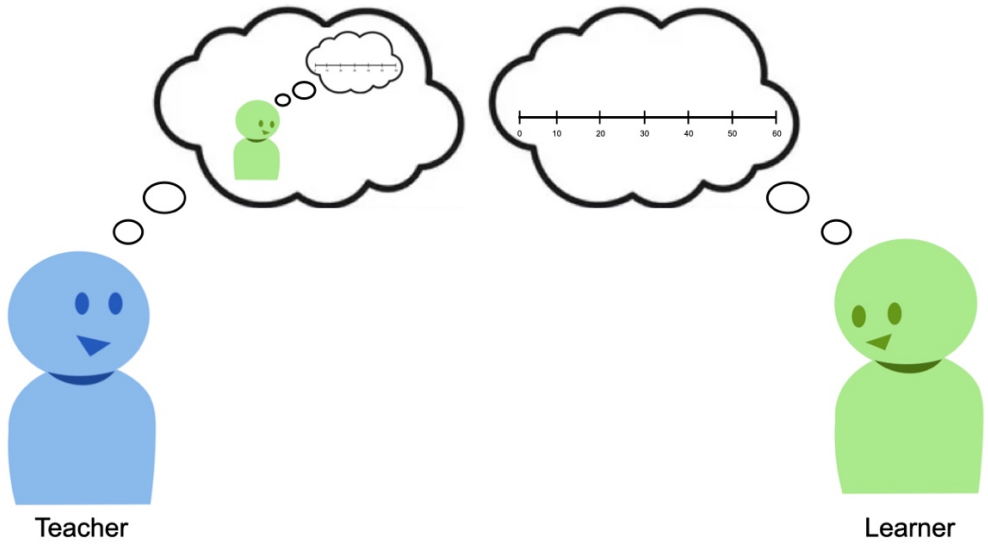

Figure 4.6: A schematic depiction of a teacher using the belief-directed strategy to provide feedback.

At the first level of the cognitive hierarchy, we model the behavior of a learner, receiving feedback directly from the line. We assume that the learner is attempting to infer the underlying reward distribution R and doing so using Gaussian process regression (GPR). GPR is a powerful, non-parametric Bayesian machine learning method that can be used to model the relationship between input variables X: $[x_1, x_2, \ldots, x_5]$ and output variables Y: $[y_1, y_2, \ldots, y_5]$ by assuming a prior distribution over the possible functions $f(x)$ that could describe the relationship and updating that distribution based on observed data. Such that:

$$y = f(x) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0,1) \tag{4.5}$$

A GP, that is, the distribution over possible functions $f(x)$ is completely defined by a mean $\mu(x)$, and a covariance function, also called a kernel $k(x, x')$:

$$f(x) \sim GP(\mu(x), k(x, x')) \tag{4.6}$$

The mean function specifies the expected value of the GP at any point x (Eq. (4.7), while the covariance function specifies how correlated rewards are between inputs x and x' (Eq. (4.8)).

$$\mu(x) = \mathbb{E}[f(x)] \tag{4.7}$$

$$k(x, x') = \mathbb{E}[(f(x) - \mu(x))((f(x') - \mu(x'))] \quad (4.8)$$

For our learner, the input variables are the chosen numbers from the line, and the output variables are the feedback values received. The learner uses those as inputs to model the relationship between chosen numbers and feedback values using GPR. This gives a likely function connecting the inputs and outputs which the learner can use to predict expected feedback for any number on the line, even if it has not been chosen before. This allows the learner to generalize their knowledge and make informed decisions to maximize expected reward. Growing research shows that GPR can successfully capture people's search behavior in various environments (Lucas et al., 2015; Wu et al., 2017).

Returning to the specifics of the present experiment, we assume that once a learner selects a number x_t and receives some reward r_t, they update their posterior belief about the underlying reward function using Gaussian process regression, giving them a new prediction for expected reward at each number on the line $V(x)$. These updated beliefs are then transformed into choice probabilities using the softmax transformation:

$$P(x_{t+1}) = \frac{\exp(V(x) * \beta_L)}{\sum_x \exp\left(V(x) * \beta_L\right)} \quad (4.9)$$

Where β_L is the leaner's inverse temperature parameter.

At the second level of the cognitive hierarchy, is the teacher, who we assume is modeling a level 1 learner. The teacher's goal is to maximize the probability of the leaner selecting the target number on the next trial. Specifically, the teacher wants to maximize:

$$P(x_{t+1} = G^* | x_1 \dots x_t, y_1 \dots y_t) \quad (4.10)$$

We assume that the teacher achieves this goal, by inferring what the learner's likely estimate of the target number would be, for each possible feedback the teacher could provide. More

concretely, say the learner selects 5 on the first trial ($x_1 = 5$). The teacher then decides which of the five possible evaluative feedback they should give by calculating the probability of the learner selecting the target number on the next trials given the feedback:

$$P(x_2 = G^* | x_1 = 5, y_1 = y_t) = \frac{\exp\left(V(x_2 = G^*) * \beta_L\right)}{\sum_x \exp\left(V(x) * \beta_L\right)} \quad (4.11)$$

For all values of $y_t \in (-2, -1, 0, 1, 2)$

This process gives the teacher 5 estimates $V(y)$ for how likely the learner is to infer the identity of the target correctly, if given each of the possible feedbacks. For instance, $V(y_1)$ is the probability of the learner selecting the target number on trial 2, if their selection for trial 1 ($x_1 = 5$) is given feedback of -2 ($y_1 = -2$). We assume that the teacher selects the feedback they want to give by converting these estimates into choice probabilities using the softmax transformation. So:

$$P(y_t) = \frac{\exp(V(y) * \beta_T)}{\sum_y \exp\left(V(y) * \beta_T\right)} \quad (4.12)$$

Where β_T is the teacher's inverse temperature parameter.

A teacher using this belief-directed strategy will structure their feedback by estimating $P(y_t)$ for each trial, which in turn is based on the numbers the learner has selected and the feedback the learner has received. Structuring their feedback in this manner relies on the teacher iteratively updating their inference about the learner's likely belief regarding the target number. Feedback will thus be sensitive to not just the distance between a number selected and the target number but also, the order in which numbers are selected. This in turn implies that a teacher relying on this strategy would treat the two leaners, Learner 1 and Learner 2, differently – since the teacher would draw a different inference for a learner who selects the number 38 after 37 compared with a learner who selects 37 and then 32.

To generate exact predictions, we simulated data from 500 teachers – 250 teachers per learner type (to match the actual sample size). We specified a prior mean of 0 and chose a radial basis (squared exponential) kernel function which assumes that points close together on the line share nearby reward values while points farther away are less correlated Eq. 4.13.

$$\kappa_{RBF}(x, x') = \exp\left(-\frac{||x - x'||^2}{2p^2}\right) \tag{4.13}$$

Where $||x - x'||^2$ is the squared Euclidean distance between the input vectors and p is the length scale hyperparameter that determines the smoothness of the resulting function. Smaller values of p^2 result in smoother functions, while larger values result in more wiggly or oscillatory functions. We optimized the hyperparameter by fitting it to our data to estimate the minimal value of p^2 that would ensure that the resulting posterior covariance matrix does not contain overly negative eigenvalues (i.e., variance > -0.5). The resulting posterior estimates y_t for all $x_i \in (0, 1, 2 \ldots 58, 59, 60)$ remain within the range $[-3, 3]$.

Modeling teacher feedback in this way gave us two distinct predicted patterns of feedback for each learner (Figure 4.7). According to the model, a teacher who interacts with Learner 1 (Figure 4.7, left panel) should provide strong negative feedback on trial 4. This is because a learner who selects 38 after receiving highly positive feedback for selecting 37 likely believes, albeit incorrectly, that the target number is slightly greater than 37. The model predicts that negative feedback from the teacher would help signal the incorrectness of this belief. In contrast, the model predicts that the teacher should provide more positive feedback to Learner 2 on trial 4 (Figure 4.7, right panel). This is because a learner who receives positive feedback for selecting 37 and then selects 32 maybe considering numbers within that range as the target numbers. By providing positive feedback for the selection, teachers can encourage that belief.

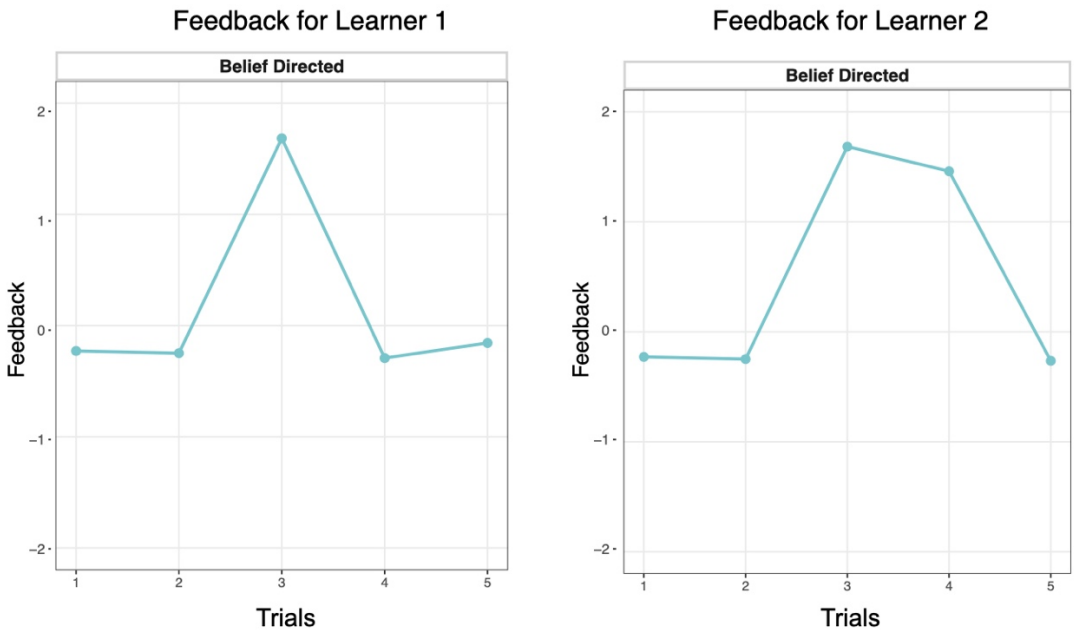

Figure 4.7: Predicted teacher feedback for Learner 1 and Learner 2 on the belief directed strategy.

4.2.3 Results

Participants interacted with only one of the two learners for 5 trials. For each learner type, we calculated mean feedback per trial and qualitatively analyzed each strategy's consistency with the participant behavior (Palminteri et al., 2017). The results, presented in Figure 4.8, suggest that the belief-directed strategy best captures the pattern observed in participant feedback compared to the three non-mentalizing strategies of mirroring, monotonic gradient, and relative gradient.

The mirroring strategy (panels A & F) and the monotonic gradient (panels B & G) both assume that teacher would treat both learners equally. The empirical data however (panels E & J) contradicts this assumption.

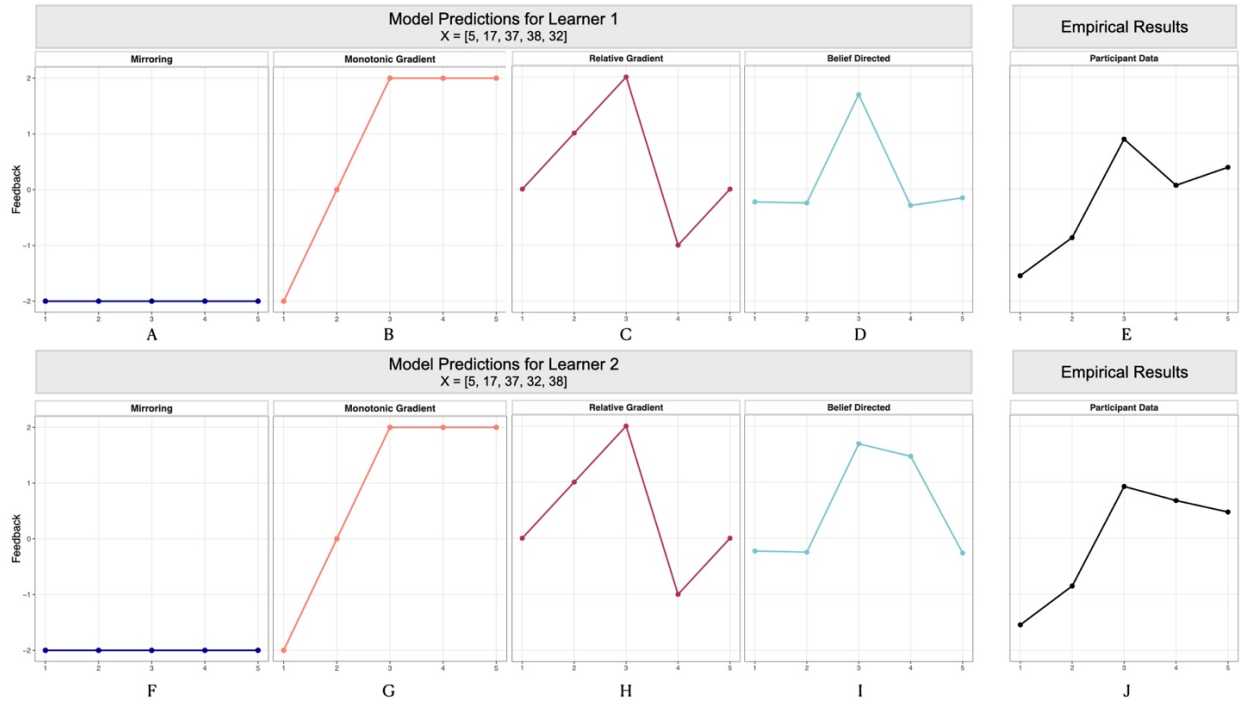

Figure 4.8: Predicted feedback for Learners 1 and 2 across various teaching strategies and data from participants. Panels A and F show the mirroring strategy's predicted feedback for Learner 1 and 2, respectively. Panels B and G depict predicted feedback from the monotonic gradient strategy for Learner 1 and 2, respectively. Panels C & H show feedback expected on the relative gradient strategy for Learner 1 and 2, respectively. Panels D & I depict feedback predicted by the belief-directed strategy for Learner 1 and 2, respectively. Finally, Panels E and J depict average feedback given by participants in Experiment 1 for Learners 1 and 2, respectively.

The relative gradient strategy, represented in panels C and H, also predicts that both learners will receive similar treatment from teachers. This prediction appears to be consistent with the qualitative pattern of feedback from teachers interacting with Learner 1. Although the model's predictions are sharper than participant data, it nonetheless captures the general pattern of the feedback. The same predicted feedback however fails to capture participant data for Learner 2.

In contrast to the three non-mentalizing models, the belief-directed strategy (panels D and I) predicts that the two learners would receive different patterns of feedback from the teachers. Participant data supports this prediction, revealing qualitatively different results for Learner 1 and Learner 2. The difference in feedback patterns presumably arises from differences in teachers' inferences about the learners' estimates of the target number.

Furthermore, the belief-directed model not only predicts that teachers would treat the two learners differently, but also captures the broad qualitative pattern of feedback both within and across the learner types. Specifically, for Learner 1, the model predicts that feedback would increase from trial 1 to 3, decrease on trial 4, and then increase again on trial 5. Participant behavior reflects this general pattern. The difference in average participant feedback from trial 4 to 5 is much less steep than the model predicts, although it is significantly different than zero ($t(236) = -5.54$, $p < .001$). Similarly for Learner 2, difference between feedback for trial 4 and feedback for trial 5 is significantly different from zero ($t(234) = 4.30$, $p < .001$). Finally, the model predicts that the average feedback on trial 4 will be lower for Learner 1 than for Learner 2. Participant data bears out this prediction. The average teacher for Learner 1 on trial 4 ($b = 0.07$) is significantly lower than average feedback for Learner 2 on trial 4 ($b = 0.60$, $t(470) = 7.50$, $p < .001$).

4.2.4 Discussion

We wanted to understand what cognitive strategy teachers employ when teaching only with evaluative feedback. To this end, we constructed a number line experiment and considered four cognitive strategies teachers could use. Three of these strategies, namely, mirroring, monotonic gradient, and relative gradient, suggest that teachers structure their feedback by picking a rule (e.g., mirror the reward distribution of the line) or a heuristic (e.g., give more positive feedback if the number selected is closer to the target and less positive feedback if the number selected is further away, etc.). On these approaches, teachers do not need to model the mind of the learner and how the learner might interpret their feedback. They need to simply pick a rule and implement it. In contrast, the fourth strategy belief-directed, predicts that teachers structure their evaluative feedback to influence the learner's beliefs to help them uncover the target number's

identity. This process requires the teacher to iteratively update their belief about the learner's likely belief about the target number. Our results, overall, suggest that people tend to structure their evaluative feedback to influence the learner's beliefs.

The belief-directed strategy seems to capture the qualitative pattern of the data most accurately across both types of learners. For both, the model predicts increasingly positive feedback for trials 1-3, as learners move from selecting numbers 5 to 17 and then 37 respectively. The model also predicts that feedback will be most positive at 37. Both predictions are easy enough to understand and could be understood as the teacher using the feedback to encourage the learner's beliefs about the target number by inferring those beliefs from the learner's number selections. On trial 4, the belief-directed strategy predicts that teacher feedback for Learner 1 should see a sharp decline. This is so because according to the model a learner picking 38 after receiving the most positive feedback for picking 37 (Figure 4.3) likely believes that the target number is near to but greater than 37. This belief is naturally incorrect as the target number is 35. A belief-directed teacher wanting to correct the learner's beliefs would therefore provide strong negative feedback to the learner for their selection of 38, thus encouraging the learner to explore numbers near but lesser than 37. Participants feedback reflects a similar pattern. The model also predicts that teachers would provide Learner 1 with relatively more positive feedback on trial 5 for selecting the number 32. This feedback is predicted to be more positive than that for trial 4 (i.e., when the learner picks 38) but less positive that than for trial 3 (i.e., when the learner selects 37). By structuring the feedback in this way, the model creates a peak in feedback around the number 37 that declines sharply for numbers after 37 and declines somewhat more gradually for the numbers between 32 and 37, making it likely that the learner would subsequently explore numbers in that region, with a preference for numbers closest to 37.

Again, participants provide feedback in a qualitatively similar manner. For Learner 2, the model predicts that teachers would give the most positive feedback for trail 3 (i.e., for selecting the number 37) followed by that for trail 4 (i.e., for selecting the number 32). Feedback for trial 5 will be much less positive in comparison (i.e., for selecting the number 38). The intuition behind the model's prediction is similar to the one before – it assuming that teachers infer the learner's likely belief about the target based on their pattern of selections. When the learner selects 37 followed by 32, they are likely considering the numbers in that region to be the target number. Therefore, providing positive feedback for 32 will elevate the set of numbers between 32-37 in the leaner's consideration. Further, by making feedback for 32 relatively less positive than that for 37, numbers closer to 37 will seem most attractive. Subsequently, selecting less positive evaluative feedback for the selection of 38 will likely demote numbers greater than 37 in the learner's consideration of the target number. Participants data seems consistent with reasoning of this kind since the pattern of feedback obtained from our participants matches the qualitative trend the belief-directed strategy predicts.

One noticeable difference between the model's predictions and data from participants is the level of smoothness in the feedback across trials. Specifically, model predictions for feedback are much sharper than people's feedback tends to be. For instance, the model predicts a steep decline in feedback when Learner 1 selects 38 after previously selecting 37. While people's feedback follows that pattern qualitatively, the differences between feedback are much smaller, leading to a smoother shift. Similarly, for Learner 2, the model predicts a small reduction in feedback from trial 3 to trail 4 but a sharp drop in feedback for trial 5. Again, participant data follows a similar trend, but it is much smoother. In fact, participants seem to treat Learner 2's selection of the number 38 much more positively than the model predicts. Interestingly, the mean

feedback that participants give to Learner 2 for selecting 38 is higher than the mean feedback they give to Learner 1 for selecting the same number (although the model predicts similar feedback for 38 both times). Why might that be the case? One simple explanation is that it could be due to measurement error. However, a different explanation is that participants seem to treat Learner 2's selection of the number 38 differently than they do Learner 1's selection. Specifically, when Learner 1 selects 38, right after selecting 37, participants might infer that the learner is entertaining numbers greater than 37 as potential target numbers. However, when Learner 2 selects 38 after selecting 37 and then 32, participants might view that selection as a way for the learner to confirm the upper boundary on the numbers they should be considering. An inference of this kind would prompt participants to respond to Learner 2 with more positive feedback than they would to Learner 1, because by doing so, the teachers can re-affirm that numbers in a nearby area are good and encourage the learners to explore locally. Therefore, although participants appear to be using a belief-directed strategy, the difference in the smoothness of their feedback compared to that predicted by the model, might come from the particular way in which people aim to influence a learner's beliefs.

The belief-directed strategy suggests that participants provide feedback by modeling the learner's reasoning and using the feedback to influence the learners' beliefs. The most efficient way for teachers to do that is to maximize the learner's informational gain on each trial through their feedback. However, a challenge with this approach is that teachers lack certainty about the learner's inference and cognitive ability, making it difficult for them to effectively shape the learner's beliefs without a clear understanding of the learner's knowledge and thought process. The teacher thus must rely on limited evaluative feedback to influence the learner's belief under a significant degree of uncertainty.

To combat this, teachers might prioritize establishing common knowledge with the learner through their feedback and then use it to impart information incrementally. A strategy of this kind would likely predict smoother feedback in ways that participants seem to provide. Recent research also seems to show that people might teach in this way. In one experiment where teachers were challenged with balancing the need to ask questions and provide instructions, they tended to ask more questions than an optimal model predicted (Popp & Gureckis, 2020) presumably to minimize their uncertainty (Bradac, 2001; Epstein, 1999).

To test whether in teaching with evaluative feedback people might employ a belief-directed strategy that prioritizes the establishment of common knowledge, we designed the next experiment.

4.3 Experiment 2

In Experiment 2, we wanted to examine the different ways in which teachers influence the beliefs of a learner when teaching a new concept. Specifically, we wanted to compare two potential approaches. According to the first, teachers shape the learners' beliefs efficiently, by maximizing the learner's information gain on each round. Abstractly, this would imply rewarding choices that provide the learner with new information about the target concept – the more information, the better. According to the second approach, teachers shape the learners' beliefs by prioritizing establishing a common ground with the learner. Abstractly, this would imply rewarding choices that make incremental (rather than maximal) informational gain and punishing choices that disregard the learner's and teacher's joint history.

To test this, we created a set of complex objects. Each object was made up of six features (shape, color, pattern, center, tone, and boundary) that could take on one of two possible values, yielding 64 unique configurations. Each object had a reward of 1 point, except one object which

had a reward of 100 points and served as the target concept (thus having the same underlying reward distribution as Experiment 1). Participants played the role of teachers. They knew the identity of the target object and the underlying reward function across the object space. Their task was to assist the learners in identifying the target object solely through evaluative feedback. Like in Experiment 1, we programmed learners to make predetermined selections to allow us to compare the predictions of each of the possible strategies a teacher could employ.

In addition to the two specific belief-directed strategies, we also consider qualitative predictions from the non-mentalizing accounts.

4.3.1 Methods

Participants 245 adult residents of the U.S. were recruited from Amazon's Mechanical Turk to take part in the experiment. After removing those who left the study incomplete, the final sample was made up of 243 participants ($M_{agegroup}$ = 25-44, 44% female identifying).

Design & materials To understand how teachers structure their rewards and punishments to help a naïve learner learn a target concept, we created an experiment involving 64 complex objects. Each object was made up of 6 features, each of which could take on one of two levels - color (red or purple), shape (square or squiggle), boundary (gray or plain), pattern (solid or stripes), tone (one-tone or two-tone), and center (empty or filled). This gave us a total of 64 unique objects (see Figure 4.9).

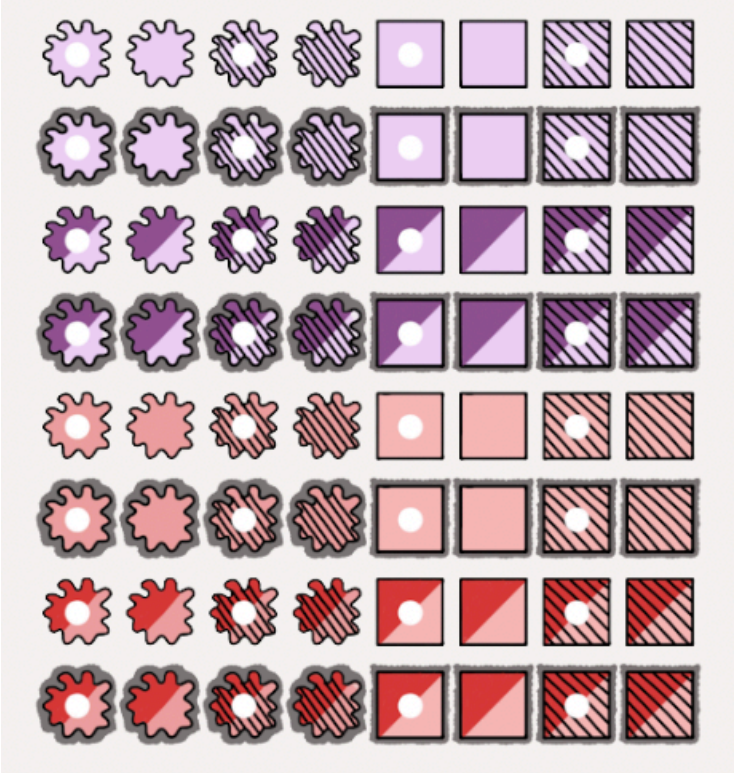

Figure 4.9: Complex object stimuli used in Experiment 2.

One of these objects was randomly selected to be the target object and assigned a value of 100 points. All other objects were assigned a value of 1 point each. Participants, who assumed the role of teachers, were informed that their task was to maximize the number of points earned. The most efficient means of achieving this was to select the target stimulus. However, teachers were not allowed to make selections directly; instead, they be paired with "learners", who teachers were told could only see the set of stimuli on their screen and not the associated rewards. The task of the teacher, therefore, was to assist the learner in earning the highest possible number of points, presumably by identifying the target concept. On each trial, teachers were presented with the stimulus chosen by the learner and the target stimulus and were required to provide evaluative feedback in the form of very bad, bad, neutral, good, or very good using the scale shown in Figure 4.2, for a total of 14 trials.

To create conditions that would allow for comparisons of the different teaching strategies, we programmed three distinct learner behaviors. All learners behaved the same in the first three trials of the experiment. Specifically, on trial 1, learners chose an object that had nothing in common with the target. On trial 2, they chose an object with exactly one feature in common with the target (always shape) and on trial 3, they chose an object with two features in common with the target (always shape & color). Then, on trial 4, all three learners made different selections. The *novel learner* chose an object that shared two new, yet untested feature-settings with the target (e.g., pattern and tone). The *incidental learner* selected an object that shared one new feature with the target (e.g., pattern) and a feature they happened to get right on just the previous trial (i.e., color). Finally, the *base learner* selected an object with one new feature (e.g., pattern) and a feature they had systematically gotten correct in the last two trials (i.e., shape) (see Figure 4.10). It is worth noting that all the three learners always selected objects that shared two features with the target. The distinction in their selections lay only in the specific features they identified correctly, with some selections consisting of previously tested and acknowledged features, while others comprising of new, untested features.

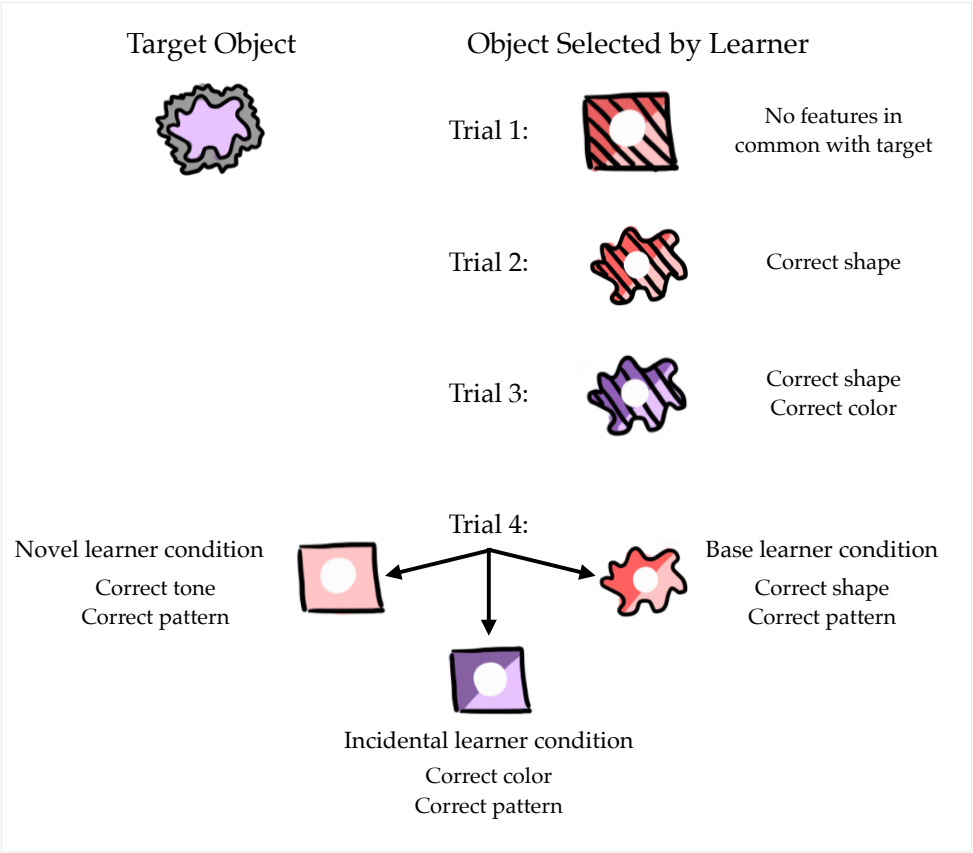

Figure 4.10: Visual representation of objects selected by the three types of learners.

Given the experimental design, we can now draw out qualitative predictions for teacher feedback using the various strategies. First, we consider the *mirroring* strategy. A teacher using this approach should provide the same feedback to all the three learners on trial 4. This is because the object selected by each learner falls outside the target object by the same number of features (i.e., 4) and is worth the same number of points (i.e., 1). So, regardless of whether the teacher is mirroring their feedback with the reward structure of the environment or the conceptual boundary between the target and non-target objects, their feedback should remain unchanged across different learners. Second, we consider the *monotonic gradient* strategy, which bases the feedback on the learner's proximity to the target in feature space. Since all three learners are at the exact same distance from the target object, in terms of the number so features

they share, this strategy predicts that all three learners should be treated the same. The r*elative gradient* strategy makes a similar prediction. Specifically, it predicts that feedback for all three learners is exactly the same since each learner's relative movement and current distance from the target are identical. For both, the monotonic gradient strategy and the relative gradient strategy, what determines the feedback is the chosen object's distance (absolute or relative) from the target in feature space, irrespective of the type of features selected. Since all three learners move the same relative distance in feature space and pick objects that share exactly 2 features with the target they should receive the same feedback. Finally, a teacher relying on the *belief-directed* strategy would treat the three learners differently. This is because the feedback from this teacher depends on their inference about the hypotheses the learner is entertaining, based on the learners' selections. As each of the three learners exhibits different patterns of selections, a belief-directed teacher would draw different inferences for each learner, thereby resulting in different feedback aimed at altering their beliefs towards the target concept.

Furthermore, our *belief-directed* teacher could attain this objective in two potential ways. One approach would be to *maximize informational gain* for the learner on every trial. This would entail rewarding the learner whenever they accurately associate a feature-setting with the target, and more so when they correctly identify new features. A teacher following the informational gain strategy would thus offer the most positive feedback to the novel learner because of the learner's correct identification of feature settings for two entirely new features, followed by the incidental and base learners. This strategy is the most efficient since it maximizes the information the learner has about the target setting for the different features.

Yet, recent research suggests that people may adopt a different approach in practice. In one such experiment, teachers were asked to instruct learners on how to assemble a tower using three

colored blocks in a specific order (Popp & Gureckis, 2020). Teachers were unaware of the learner's starting block arrangement and on each trial could either ask them a question (to gain understanding of the leaner's arrangement) or give them an instruction (to help them build the target tower). Results showed that in general, teachers asked more questions than was necessary possibly due to a bias of wanting to establish what the learner knew and provide instructions accordingly (to correct or build on the learner's block arrangement). Furthermore, teachers also electively asked more questions when they were uncertain about their learner or inferred that the learner was confused based on the learner's poor performance. Results from this experiment thus suggest that teachers might structure their evaluative feedback to influence the beliefs of a learner in a manner that prioritizes *establishing common ground* with the learner and using that to successively impart new information. One way for the teacher to do this would be to selectively reward a specific feature (e.g., shape) each time the learner picks the correct setting. Subsequently, once the teacher is confident that the learner understands the correct setting of that feature, they may start to reward a second feature selectively – hoping that the learner retains the formerly established knowledge. A teacher following this approach in our experiment would accordingly give the most positive feedback to the base learner and the most negative feedback to the novel learner. This is because, the base learner builds on common knowledge established by the teacher by retaining a previously rewarded feature (e.g., shape) and testing one new feature (e.g., pattern). In contrast, the novel learner disregards the previously established common knowledge. This learner correctly identifies two new features (maximizing their informational gain) but does so at the expense of foregoing gains they established through collaborating with the teacher, thus earning them the most negative feedback.

4.3.2 Results

To examine the relationship between participant feedback and different types of learners, we transformed the evaluative feedback scale into an ordinal scale, with a range of -2 to 2. We then regressed participants feedback on trial 4 (the critical trial) on the different learner types using a linear model. To account for the various possible hypotheses, we created two dummy codes, one for the base learner and another for the novel learner, making incidental learner our reference category. Results reveal that the model fit was significant ($F(2, 240) = 24.91$, $p < .001$). Overall, all three learners received negative feedback (see Figure 4.11). Compared to our reference category, the incidental learners ($M = -0.93$, $b = -0.93$, $p < .001$), base leaners received more positive feedback ($M = -0.27$, $b = 0.65$, $p < .001$) while novel learners received more negative feedback ($M = -1.32$, $b = -0.4$, $p = .009$). These results fall in line with the qualitative predictions of a teacher structing their feedback to shape the learner's belief using a strategy to minimize uncertainty by establishing common ground.

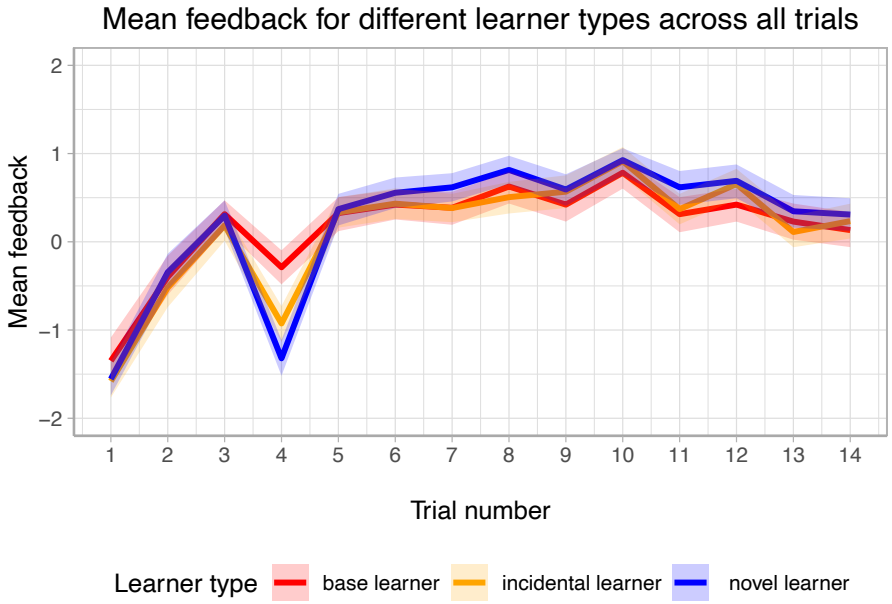

Figure 4.11: Mean feedback for each learner type.

4.3.3 Discussion

In experiment 2 we wanted to test the various strategies a teacher could employ to structure their evaluative feedback when teaching a new concept. We further wanted to compare two potential approaches teachers could use to shape a learner's beliefs – maximizing the learner's information gain on each round or establishing common ground with the learner. To this end, we created three learners that on the fourth trial made selections that would lead them to be treated differently by each of the various approaches. The base learner selected an object that shared one new feature with the target (pattern), along with a feature that they had correctly identified in two prior trials (shape) – implying that they were utilizing the knowledge established between themselves and the teacher. The incidental learner chose an object that shared one feature with the target (pattern) and a feature they correctly identified in just the previous trial. Finally, the novel learner selected an object that correctly identified two entirely new features with the target object – maximizing the informational gain but at the cost of giving up shared common ground. Results showed that of all available strategies – mentalizing and non-mentalizing, participant behavior matches predictions of the mentalizing, belief-directed strategy most accurately as participants treat the three learners differently. Furthermore, the specific pattern of difference is illuminating. In general, participants give all three learners negative feedback. Specifically, they give the most negative feedback to the novel learner and the least negative feedback to the base learner. These results therefore suggest that when trying to shape the beliefs of a learner using evaluative feedback, teachers tend to prioritize establishing common ground between themselves and the learner, instead of maximizing informational gain.

4.4 General Discussion

We explore how people teach intuitively using rewards and punishments. In our experiments, participants, playing as teachers, were shown target concepts (number line or a set

of complex objects), and tasked with helping learners identify the target concept using only evaluative feedback.

To understand how teachers structure their feedback, we examined four potential strategies: mirroring the environment, utilizing distance-based heuristics such as monotonic gradient or relative gradient, or engaging in a form belief-directed pedagogy, using an approach to either maximize informational gain through feedback or establish common ground through feedback. The results from both experiments demonstrate that neither mirroring nor heuristic strategies are employed alone. For example, in Experiment 1, teachers give varying feedback across trials suggesting that mirroring alone cannot explain the feedback since each number has the same reward. Similarly, in Experiment 2 as the learner progresses from correctly identifying no features to correctly identifying one feature, the average feedback improves, even though the underlying reward function is unchanged in this region. Furthermore, a simple heuristic strategy, which focuses on minimizing the distance between the current selection and the target (in number space or in feature space), also seems insufficient since in Experiment 1 participants treat the learner's selections of the numbers 32 and 38 differently across trials even though both their distances from the target are the same. Similarly, on the critical trial in Experiment 2 all three types of learners—the base learner, incidental learner, and novel learner— receive different feedback from the participants, even though each learner selects an object that shares exactly two correct features with the target, following a trial on which their choice shared two features.

Naturally, there may be additional non-mentalizing strategies beyond the ones we have considered. In fact, it is possible to devise strategies that accurately explain teacher behavior without reference to mentalizing. However, while tailored strategies of this kind may explain specific teacher behaviors under certain conditions, they may lack generalizability. Our goal, in

contrast, is to identify general principles underlying how teachers use evaluative feedback, such as punishments and rewards.

As such, our findings suggest that teachers tend to employ principles of belief-directed reasoning, utilizing feedback strategically to influence the learner's understanding given a model of the learning process. Additionally, the pattern of results also provides insight into the nature and structure of this pedagogy. A teacher could aim to shape a learner's belief in at least two ways: by maximizing informational gain across all dimensions simultaneously for each trial, or by reducing their own uncertainty about the learner's inferences by establishing common ground and then imparting information successively. Our results provide evidence for the latter approach. We find, in Experiment 2, that a novel learner receives the most negative feedback as they give up the two features they had previously gained with the teacher.

Our findings may seem surprising at first—after all, the teacher's strategy of minimizing their own uncertainty is suboptimal compared to the strategy that prioritizes informational gain for the learner. Nevertheless, recent studies suggest that human reasoning may, in fact, exhibit suboptimal behavior along similar lines (Popp & Gureckis, 2020). It is also worth noting that the complexity of the stimuli we used in the experiment, which consisted of six different features, may have contributed to the adoption of the cognitively simpler strategy of establishing common ground. It's therefore an open question whether people's adoption of this strategy is specific to the current experimental design or if it reflects a more general tendency of how people teach with evaluative feedback. To that end, a fruitful next step would be to simplify the feature space and see if there is a shift in teacher strategy.

Bearing this caveat in mind, our results suggest that people may structure their evaluative feedback communicatively, using their capacity for recursive theory of mind. On this view,

teachers have a mental representation of the learning process (i.e., their generative model), and they use this to identify the specific evaluative feedback that will maximize the probability of the learner choosing the target concept on the next round. Meanwhile, we propose that learners likely recognize the teacher's pedagogical communicative intent and learn from the feedback accordingly (although learner behavior was not directly explored in our experiments). Mutual inference of this kind can be best modeled as recursive Bayesian inference. While we have taken a first step in building a formal computational model for this type of pedagogical interaction, further modifications are necessary, particularly to account for the teacher's tendency to prioritize establishing common ground when using a belief-directed strategy, as observed in Experiment 2. This area remains an important topic for future research.

Altogether, our study provides a valuable initial exploration into how teachers teach with evaluative feedback in instances of learner driven pedagogy. However, there are a couple of caveats. First, we manipulated the behavior of the learners in both experiments to control for the choices they made. For instance, in Experiment 1 we programed the learner to select a very specific set of numbers which made it easier to differentiate between teaching strategies. However, this leaves open the question of how teachers would respond to learners making different selection of numbers across the line. Future research could test the predictions of these strategies against various learner behaviors, including real learner behavior. Similarly, in Experiment 2 learners were programmed to select the same correct features on trials 1-3. Specifically, all learners selected shape as the base feature and color as the incidental feature. This enabled us to tightly control the behavior of the learners and measure differences solely in teacher response. However, features such as color and shape may be more salient and easier to communicate about than features such as pattern or tone. Therefore, future work must expand

upon these findings by utilizing different features as the base and incidental features. A second caveat to bear in mind is that we programmed the learners in the experiment to make selections in a way that made it easy to distinguish between different strategies. However, in real life, people likely use a mix of simple, heuristics strategies and more complex mental state inference strategies when giving feedback. Finally, a third caveat is that both experiments employed a 'needle-in-the-haystack'-like reward function, according to which, the target concept (i.e., target number or object) was worth 100 points while all other choices (i.e., other numbers or objects) were worth only 1 point. We adopted this reward function to simplify the experimental set up and control for confounding factors but in future work it would be beneficial to introduce a varied reward distribution. This would allow us to observe how teachers may trade-off information, such as closeness in distance or feature space with low or negative reward.

5

Conclusion

Of the many ways in which humans influence the behavior of others, a particularly potent one is through punishment. Punishment is a powerful tool that can be used for correcting misaligned values of individuals who committed moral infractions, and for instilling new values, to those acquiring new knowledge. We have a natural inclination to punish (Boyd & Richerson, 1992; Fehr & Gächter, 2002; Morris et al., 2017), which emerges early in life (Bernhard et al., 2020; Cushman et al., 2013) and is observed across various cultures (Boehm, 2011; Henrich, 2006; Henrich et al., 2006) and a complimentary remarkable ability to quickly learn from the punishments imposed on us (Azrin & Holz, 1966; Fehr & Gächter, 2002; Guitart-Masip et al., 2012; Thorndike, 1927). In this dissertation I attempt to elucidate the cognitive mechanism that make this possible. Specifically, I examine the cognitive capacities and constrains that shape the contours of how we teach with and learn from punishments.

To this end, I consider two possibilities. The first possibility is that we engage with punishment by treating it as a form of reinforcement, which relies on our cognitive capacity for reinforcement learning. As humans, we possess an inherent inclination to modify our behavior in response to the outcomes they produce. This tendency leads us to avoid behaviors that result in

unpleasant consequences. On this view, punishment is effective in modifying subsequent behavior because of the aversiveness of the punitive act itself. For example, when a child receives a time-out for pushing a friend on the playground, their behavior improves because the time-out is an aversive experience. Engaging with punishments in this way is simple and requires punishers to impose costs and for recipients of respond to those costs. The second possibility is that we teach with and learn from punishments by treating them as a form of communication, which relies on our cognitive capacity for theory of mind. As social beings, we are constantly surrounded by other people, and we often make sense of their behavior by attributing underlying beliefs, desires, and knowledge to them. The attributions and inferences we make play a crucial role in shaping our perceptions of people and the subsequent decisions we make. Viewed through this lens, punishment is an informative signal that communicates something about the punisher's values and goals. What makes this signal effective in modifying subsequent behavior is the message it conveys and the inferences it enables others to make. Thus, the reason a time-out results in better playground behavior is because it signals the parent's disapproval of the child's misconduct, and it is this message that encourages the child to change their behavior. On this view, punisher and recipient are coordinated in expressing and inferring communicative intent through the punishment.

While these possibilities are distinct, they are neither mutually exclusive, nor can they fully explain our engagement with punishment on their own. Typical punishments impose both a cost and communicate a message, allowing individuals to learn by relying on either mechanism. In some instances, punishments are imposed immediately after a violation. They are timely and topically related to the violation, allowing simple principles of reinforcement learning to capture the learning process. However, at other times, punishments may come later and may not impose

direct costs but may nonetheless signal social disapproval or a misalignment of one's values with those of the group. To accurately identify and learn from such punishments, individuals would need their theory of mind to understand the social context of the punishment and infer its likely reasons. Furthermore, while people likely pay attention to both the cost of a punishment and the communicative message, it's possible for the punisher and recipient to be misaligned in their coordination. For instance, punishers could structure their feedback to be maximally communicative, while recipients could be focused solely on the imposed incentives. Therefore, identifying the distinct roles of each mechanism and understanding how they work together is crucial for comprehensively understanding how we teach with and learn from punishments.

The research shared in this dissertation suggests that individuals generally conceive of punishments as organized around principles of communicative inferences, in both moralistic and non-moralistic situations. However, there are instances where the direct reinforcement value of punishment is crucial in bringing about the desired behavioral change. Specifically, Chapter 1 shows that people typically engage with punishment through communicative inference, where teaching and learning involve expressing and inferring communicative intent. Chapters 2 and 3 identify situations where the direct unpleasantness of punishment plays a pivotal role in eliciting the desired behavioral change. Finally, Chapter 4 demonstrates that when people use punishment to impart information to teach new concepts, they structure it communicatively to influence the learner’s beliefs about the target concepts rather than to simply reinforce good or bad actions. Collectively, this work highlights the cognitive capacities underlying our ability to engage with punishment and contributes to the extensive literature on the psychology of punishment.

5.1 Advancements to literature & directions for the future

CONCLUSION

Chapter 1 demonstrates that individuals' usual approach to punishment is centered around principles of communicative inference. Accordingly, punishers use the communicative signal of the punitive act to express their message and recipients learn from punishments by inferring the intended message. In doing so, this work makes two important contributions. First, it sheds light on the source of punishments' efficacy. It shows that contrary to previous assumptions, the aversiveness of punishments does not derive solely from the punitive act itself (Hofmann et al., 2018). Rather, the social-communicative message that punishments convey also significantly contributes to their aversiveness. Second, it places focus on the punitive interaction, demonstrating the significance of understanding punishment as a dynamic system involving different individuals. Prior work on punishment has often examined its elements in isolation, such as the motives or emotions of the punisher (Carlsmith, 2006; Funk et al., 2014; Jordan & Rand, 2020; Raihani & McAuliffe, 2012), or how the context and crime affect punishment judgments (Margoni & Surian, 2021; Martin & Cushman, 2016; Nobes et al., 2009; Shultz & Wright, 1985). While this research has given us valuable insights into the individual pieces that make up the current mosaic of research on the subject, shifting our attention to the punitive interaction will enable us to obtain a more comprehensive picture of the dynamics at play. The analogy to communication is particularly fitting here. We can understand speech and language on their own but to fully comprehend communication, we need to develop a joint understanding of the speaker and the listener. Similarly, a comprehensive understanding of punishment requires that we develop a joint understanding of both the punisher and the recipient, including the ways in which communicative intent is expressed and inferred and the implications this process has for learning outcomes.

Empirical studies provide some suggestive evidence that the degree to which recipients of punishments cooperate with their punisher is influenced by the inference they draw of their punisher's motive (Xiao, 2013). For instance, in a two-player Trust Game, trustees who coerce their partners into investing and returning a large sum by threatening punishments receive less than those who refrain from making such threats (Fehr & Rockenbach, 2003; Houser et al., 2008). The authors suggest that the differences arise from the inferences made by the partner about the trustee's underlying motives, whereby partners infer that the trustee has greedy or selfish intentions and is using sanctions to enforce unfair distribution of resources (Fehr & Rockenbach, 2003). Similarly, a meta-analysis conducted by (Balliet et al., 2011) reveals that the efficacy of reducing a co-player's payoff to foster cooperative behavior is dependent on whether the punishment is costly or cost-free for the punisher. This, the authors argue is because, when punishment is costly to administer, it more clearly conveys that the punisher has benevolent motives and is using punishment to endorse collective norms and foster cooperation, rather than merely advancing self-interest. Building on these insights, an exciting avenue for future research would be to manipulate the inferences drawn by recipients of punishments and examine the impact on their behavior. For instance, we could administer punishments to recipients and inform them that the punisher either directly benefited or did not benefit from the punishment, thereby yielding divergent inferences about the punisher's motive. We could then measure the recipients' inferences about the reasons for the punishment, their decision to cooperate immediately after the punishment, and their desire to maintain an ongoing relationship with the punisher.

A valuable complement to this would be to explore how punishers balance their goals when selecting punishments. Punishers may have multiple goals and motives to punish, including expressing disappointment, reaffirming values, seeking revenge, gaining a competitive

advantage, or educating the recipient. By treating punishment as a form of communication, we can examine how punishers might use different types and modalities of punishment to convey different messages and achieve different goals. For example, verbal punishment may convey more explicit information about the punisher's expectations and reasons for punishment, while physical punishment may convey more emotional and affective signals about the punisher's disapproval or anger. The analogy to communication is useful once again. People often use figurative speech, such as hyperbole, to convey a message beyond its literal meanings, allowing listeners to draw richer inferences about the speaker's feelings (Goodman & Frank, 2016; Kao et al., 2014). For instance, by saying this coffee cost “a million dollars” instead of saying it cost “a lot” a speaker conveys not just literal information, that coffee was expensive, but also their surprise and frustration. Thus, despite deviating from Gricean principles (Grice, 1957), figurative speech enhances communication by conveying a more nuanced message (Goodman & Stuhlmüller, 2013). Likewise, it would be interesting to investigate how punishers construct their punishments to express various motives, and how different types of punishments amplify certain inferences while diminishing others.

It would also be advantageous to extrapolate these insights to think more deeply about punishment in the real world, such as institutional punishment. Punishments of this kind are incredibly common and often carried out by individuals who represent institutions such as the government, an organization, a club, or broadly any collectively endorsed social group. What’s interesting about these contexts is that the person carrying out the punishment (e.g., a judge, a coach, a club president) often holds a position of authority and advantage over the recipient of the punishment. This marked power asymmetry can impact the inferences drawn by the recipient and, in turn, their ability to learn from the punishment. For instance, if a student arrives late to a

debate team practice and their coach decides to punish them, the coach's intentions may be just and reasonable. Nevertheless, the student may perceive biased behavior on the part of the coach and accuse them of being unfair. Our theory of mind capabilities are susceptible to limitations and biases that can result in drawing incorrect inferences. These distortions may be further amplified in situations where the punisher possesses unilateral power over the recipient. If an individual perceives the punishment as unjust, it may lead them to lose faith and trust in the institution, which may subsequently lead to a disregard for the social norms that the punishment aimed to reinforce. Therefore, identifying the factors that can corrupt cooperative inferences, notably within the context of institutional punishments, will allow us create justice systems that not only administer justice equitably but also appear unbiased and fair across individuals and over time.

In a related vein, there is often a significant degree of uncertainty involved in accurately expressing and inferring communicative intent. As punishers, we are not privy to what the learner has truly understood and inferred from the punishment. As recipients, we lack perfect knowledge of the punisher’s underlying motives. Moreover, people may sometimes engage in strategic and deliberate deception, such as pretending to comprehend the message and signaling their intent to change, even when they have no genuine desire to do so. It would therefore be beneficial to investigate how changes in objective and perceived uncertainty affect the choice of punishment and the willingness to learn from it.

Chapters 2 and 3 demonstrate how the negative incentives of punitive acts directly prompt favorable changes in the recipient's behavior. Both chapters examine cases of negligence, in which a person unwittingly causes harm by failing to bring a known and relevant thought to mind at the right time. Both chapters also make significant contributions to existing literature.

Chapter 2 shows that punishers willingly punish mere failures of thought. Drawing on insights from cognitive science, Chapter 3 proposes how the aversiveness of the punitive act can automatically alter what comes to the recipient's mind in the future, in turn promoting favorable changes in their behavior. Together they highlight the role the direct unpleasantness of a punishment plays in making it effective.

Although both chapters focus specifically on negligence, their insights can be applied to other failures of thought, such as microaggressions. Microaggressions are subtle, often unintentional comments or behaviors that communicate a derogatory or negative message towards a member of a marginalized group based on their race, gender, sexual orientation, ethnicity, or disability. Like instances of negligence, individuals often express microaggressions unintentionally or without awareness. Findings from these chapters can help us understand why people might feel compelled to punish instances of microaggressions and how such punishment could ultimately reduce the likelihood of future microaggressions.

The conceptual insight derived from this research can prompt us to reflect more critically on the types and severity of costs associated with punitive measures and their potential implications for the recipient's subsequent behavior. While punishments that are too mild may not effectively deter future bad behavior, punishments that are excessively harsh (especially in comparison to the transgression) may elicit short-term cooperation but may undermine the autonomy of the recipient, impede their internalization of the desired norm, and their propensity to comply with others over the long term (Balliet et al., 2011; Deci & Ryan, 2000). For instance, harsh punishments, such as corporal punishments, lead to short-term compliance and behavioral change, but they have negative long-term consequences, such as increased delinquency and involvement in criminal activities (Afifi et al., 2017; Gershoff, 2002; Joyner & Beaver, 2022;

Maguire-Jack et al., 2012). When an imposed punishment is perceived as an attempt for the punisher to assert their power and to control the recipient's autonomy, it is likely to lead the recipient to retaliate, sometimes by engaging in the very behavior were punished for (Brehm, 1966; Moore & Pierce, 2016). Such reactance by the recipient is contingent on their inferences of the punisher's motives and may be more likely to occur when the costs imposed are perceived as excessively high or unjust. It would thus be fascinating to explore how various types and costs of punishments influence the recipient's inferences regarding the punisher's motives (and vice versa) and the implications these have for the recipient's willingness to learn and cooperate. This work could help shed light on the appropriate balance that punishers should maintain when administering punishments that are constructive rather than damaging.

The findings from this dissertation could also provide a novel perspective in understanding instances of self-harm. Self-harm involves intentional self-inflicted actions aimed at causing injury or harm to oneself, such as cutting, skin picking, restricting food intake, etc. It is often used as a coping mechanism to deal with emotional distress or trauma. While the reasons for engaging in self-harm vary widely, a common reason is self-punishment driven by levels of self-criticism, where individuals believe that they deserve pain and punishment (Fox et al., 2017; Glassman et al., 2007; Hooley & St. Germain, 2014; Nock & Prinstein, 2004; Swannell et al., 2008). While the current clinical research on the topic highlights the key role self-punitive sentiments play in producing self-harm utilizing our understanding of why and how we punish others might offer a complimentary lens to understand reasons and triggers behind self-punishment. As this dissertation outlines, our punishments for others can be structured as forms of reinforcements or forms of communications. Extensive extant research (in non-clinical areas) also details the various granular factors (personal, situational, and cultural) that influence our

punishments for others. Might this research offer us some leverage in understanding self-punishment? We may discover that the two instances punishment – for others and for oneself – may differ and requires distinct considerations. However, we might also identify similarities that enable us to devise interventions that could predict self-harm in individuals who self-injure to punish themselves.

Finally, in Chapter 4 I examine whether rational models of pedagogy can help explain how teachers give out evaluative feedback in instances of pedagogy. I find that teachers tend to structure their feedback communicatively, dynamically influencing the learner's beliefs, rather than relying solely on simple rule-based or heuristic strategies. Furthermore, they give feedback in a way that prioritizes establishing common ground between themselves and the learner. These findings highlight another avenue where people rely on their capacity for theory of mind to engage with punishments.

A predominant focus in the study of theory of mind has been on its predictive capacity, to understand and anticipate the behavior of others. However, a recent proposal has brought to light a complementary aspect of theory of mind – Its use in action planning (Ho et al., 2022). For instance, to surprise your friend with lunch, you could hypothesize that they typically have their lunch around noon, and thus predict that they may leave their office approximately 10-15 minutes before that to grab lunch. Then, using this prediction, you could plan your arrival accordingly to catch them before they leave. The outcomes from Chapter 4 furnish initial yet compelling evidence of planning with theory of mind within the context of teaching with evaluative feedback. Our cognitive model, which draws inspiration from recursive models of Bayesian inference, shows that the teacher's first-order mental state inference, where they use

their own generative model of the learner's behavior to plan their feedback, can effectively capture the actual ways in which people provide feedback.

These empirical and conceptual insights open several interesting questions. To effectively plan their teaching processes and feedback, teachers must solve several related challenges. For instance, to teach someone to become a professional tennis player, a teacher must impart a diverse range of skills, from concrete skills, such as proper racquet holding, foot movement, and serving, to highly abstract skills like planning a game strategy and reading the opponent's moves. How do teachers break down these goals and decide on the appropriate level of abstraction for their feedback? Since if the feedback is too broad or too specific it can be counterproductive. Then, once they have decided on the teaching process, how do they organize the sub-goals hierarchically? Since progress can be hindered by feedback that is excessively complicated or disorganized. Relatedly, when do they decide to offer positive or negative feedback to the learner? How teachers approach these challenges may depend on several other factors, one of which might be the learner's cognitive capacity. For instance, one could reason that young children may benefit from concrete feedback focused on actions and behaviors, while adults may benefit from more abstract feedback highlighting the overall goal of the teaching endeavor and the teacher’s approach. Moreover, even after identifying and organizing their teaching goals, teachers may dynamically modify them during the teaching processes. Gaining a better understanding of how teachers approach and solve these challenges would be a fruitful future endeavor, given that pedagogy is a common and frequent activity. Furthermore, exploring the cognitive representations that facilitate these pedagogical interactions would also make for fascinating research.

5.2 Notion of punishment

Throughout the dissertation, I have used the term punishment to mean the imposition of cost. However, individuals can also punish by withholding or withdrawing benefits. Punishing in this way is common in the real world, especially in interpersonal situations where there are no power asymmetries and therefore a higher risk of retaliation. When people are given opportunities to punish in different ways, laboratory experiments show that people prefer to withhold opportunities (Balafoutas, Nikiforakis, et al., 2014) or distance themselves from transgressors rather than actively harming them (Boehm, 2001; R. B. Lee, 1979). However, prior research shows that withholding benefits has the same psychological impact as imposing punitive costs (Lutz & Widmer, 2014; Seymour et al., 2007; Wagner, 1969). Therefore, people can learn from punishments that involve the absence of rewards in the same way they learn from aversive costs: by paying attention to the (costly) absence of a rewarding experience or by inferring why the punisher withheld the benefit. However, how effective these punishments are compared to other types of punishments and what kind of inferences they evoke is an empirical question. For example, a non-dishwashing roommate could be punished by leaving dirty dishes on their bed, leaving a brand-new dish soap and a sponge on their bed, or by not attending a planned movie night. While all these methods communicate disapproval and elicit aversion, they may differ in their level of aversiveness, the inferences they convey about the punisher's motives and goals, and in how effective they are in bringing about a desirable change in the recipient’s behavior.

5.3 Interdependence between the cognitive mechanisms

The two possibilities I consider, for how we teach with and learn from punishments, derive from cognitive mechanisms we have. The possibility that we treat punishment as reinforcement is rooted in our capacity for reinforcement learning while engaging with punishments as

communications relies on our capacity for theory of mind reasoning. Although these are distinct cognitive systems, they often work together. As previously mentioned, a typical punishment involves both an unpleasant cost and a clear message. Thus, teaching and learning from punishment involves both systems to work together. Furthermore, these systems also work in tandem to support many other aspects of our social cognition. Bayesian inference, which formalizes theory of mind, relies on generative models which describe how beliefs, desires, and goals influence behavior. However, these models also incorporate representations of rewards and punishments based on our knowledge of and experiences with positive and negative outcomes. These representations of rewards and punishments help shape our expectations about how different mental states influence our behaviors and beliefs about the world. By inverting our own generative models, we successfully infer the mental states of others from their actions, a process referred to as inverse reinforcement learning (FeldmanHall & Nassar, 2021; Jara-Ettinger, 2019). Recent studies suggest that the theory of mind reasoning process can be modeled as inverse reinforcement learning in certain pedagogical interactions (Ho et al., 2016). Therefore, the cognitive mechanisms underlying punishment are intricately linked and enable individuals to navigate diverse and complex social interactions, including but not limited to punishment.

5.4 Observational learning

Throughout the dissertation I have focused on instances where people learn from punishments through their own direct experience with actions that lead to punishment. However, another essential way for us to learn from punishments is by observing other people perform actions and get punished for them (Bandura, 1977; Olsson & Phelps, 2004). A prime advantage of learning in this way is that it lets us learn which actions to avoid without having to do those actions or experience their negative consequences ourselves. Although for this to be successful,

the action and aversive consequence must be clearly visible and follow close in time. Otherwise, it can lead to the problem of credit assignment (Bellman, 1954). Observational learning may also play a vital role in the cultural learning of behavioral norms, across people and across generations. Consequently, it is possible that in addition to learning appropriate behaviors, people might also culturally acquire the propensity to punish as well as the manner and method of punishment deemed appropriate (Gergely & Csibra, 2005). An interesting avenue for future research would be to use evolutionary analysis and reinforcement learning simulations to understand how punitive behaviors and goals are acquired over successive generations and the implications they have for cooperation.

5.5 Conclusion

Collectively, the research presented in this dissertation illuminates the roles that our abilities to learn from reinforcements and theory of mind play in shaping our experience with punishments – how we think of them, teach with them, and learn from them. The insights gleaned from the research suggest a promising avenue for further punishment research is to examine the dynamics of punitive interactions by considering the perspectives of both the punisher and recipient. Such an approach would enable us to identify ways in which punishments can foster effective learning outcomes, as well as ways in which they may be maladaptive. This, in turn, will enable us to develop a more nuanced and comprehensive model of human punishment that can serve as a unified framework for future research.

A

Supplementary materials for Chapter 1

APPENDIX A

A.1 Vignettes used in Experiment 1

A.1.1 Dirty dishes

	Costly	*Costless*
Informative	The next day, before Sandra gets back home, Alice leaves some of Sandra's dirty dishes on her pillow with a ribbon around them and a tag that says: "Love, your roommates".	The next day, before Sandra gets back home, Alice leaves a brand-new kitchen sponge and bottle of dish soap on her pillow with a ribbon around them and a tag that says: "Love, your roommates".
Uninformative	The next day, before Sandra gets back home, Alice gathers all of Sandra's unopened mail, which was scattered around the living room. Then, Alice throws all of Sandra's mail across the floor of Sandra's room, leaving a note on her pillow that says: "Love, your roommates".	The next day, before Sandra gets back home, Alice gathers all of Sandra's unopened mail, which was scattered around the living room, and puts it into a neat pile. Then, Alice leaves the neat pile of mail on Sandra's pillow with a ribbon around it and a tag that says: "Love, your roommates".

Background text: Alice lives in an apartment with three other roommates. One of them, Sandra, has a very frustrating habit of leaving her dirty dishes in the kitchen sink. The roommates have pointed out the dishes to Sandra but they don't know exactly how to say the right thing, and she keeps doing it. Alice talks with her other roommates and together they decide to do something about it.

A.1.2 Messy mail

	Costly	*Costless*
Informative	The next day, before Sandra gets back home, Alice gathers all of Sandra's unopened mail, which was scattered around the living room. Then, Alice throws all of Sandra's mail across the floor of Sandra's room, leaving a note on her pillow that says: "Love, your roommates".	The next day, before Sandra gets back home, Alice gathers all of Sandra's unopened mail, which was scattered around the living room, and puts it into a neat pile. Then, Alice leaves the neat pile of mail on Sandra's pillow with a ribbon around it and a tag that says: "Love, your roommates".
Uninformative	The next day, before Sandra gets back home, Alice leaves some of Sandra's dirty dishes on her pillow with a ribbon around them and a tag that says: "Love, your roommates".	The next day, before Sandra gets back home, Alice leaves a brand-new kitchen sponge and bottle of dish soap on her pillow with a ribbon around them and a tag that says: "Love, your roommates".

Background text: Alice lives in an apartment with three other roommates. One of them, Sandra, has a very frustrating habit of leaving her mail unread and scattered around the living room. The roommates have pointed out the mail to Sandra but they don't know exactly how to say the right thing, and she keeps doing it. Alice talks with her other roommates and together they decide to do something about it.

A.1.3 Loud music

	Costly	*Costless*
Informative	The next day, before Scott gets into work, Adam removes the batteries from Scott's portable speakers and in their place leaves a note that says: "Love, your colleagues".	The next day, before Scott gets into work, Adam leaves a pair of cheap earphones on Scott's desk with a ribbon around them and a tag that says: "Love, your colleagues".
Uninformative	The next day, before Scott gets into work, Adam leaves a box of rotten food on Scott's desk with a ribbon around it and a tag that says: "Love, your colleagues".	The next day, before Scott gets into work, Adam leaves a new, empty lunchbox on Scott's desk with a ribbon around it and a tag that says: "Love, your colleagues".

Background text: Adam works in an office and sits on a desk which is surrounded by other desks belonging to his colleagues. One of Adam's colleagues, Scott, has a very frustrating habit of listening to music on his portable speaker whilst working. Adam and all of the neighboring colleagues can always hear the music and are unable to concentrate on their work. The colleagues have pointed out the music to Scott but they don't know how to exactly say the right thing and Scott keeps listening to music on his speakers. Adam and his colleagues decide to do something about this.

A.1.4 Lunch

	Costly	*Costless*
Informative	The next day, before Scott gets into work, Adam leaves a box of rotten food on Scott's desk with a ribbon around it and a tag that says: "Love, your colleagues".	The next day, before Scott gets into work, Adam leaves a new, empty lunchbox on Scott's desk with a ribbon around it and a tag that says: "Love, your colleagues".
Uninformative	The next day, before Scott gets into work, Adam removes the batteries from Scott's portable speakers and in their place leaves a note that says: "Love, your colleagues".	The next day, before Scott gets into work, Adam leaves a pair of cheap earphones on Scott's desk with a ribbon around them and a tag that says: "Love, your colleagues".

Background text: Adam works in an office where he and his colleagues all share a common fridge. Most employees get their respective lunches to work and leave it in the office fridge. However, one of Adam's colleagues, Scott, does not bring his own lunch to work. Instead, Scott has a very frustrating habit of eating other people's food from the common fridge. Adam and the others have pointed this out to Scott but they don't know how to exactly say the right thing and he keeps eating their food. Adam and his colleagues decide to do something about this.

A.1.5 Stationary scenario

	Costly	*Costless*
Informative	The next day before Sandra arrives at the lab, Alice covers Sandra's entire desk in sticky notes and leaves a note on the desk that says: "Love, your lab mates".	The next day before Sandra arrives at the lab, Alice leaves a giant cardboard box full of sticky notes with a ribbon around them on Sandra's desk along with a note that says: "Love, your lab mates".
Uninformative	The next day before Sandra arrives at the lab, Alice leaves a pile of sweaty and stinky clothes on Sandra's desk with a ribbon around them and a note that says: "Love, your lab mates".	The next day before Sandra arrives at the lab, Alice leaves a brand-new laundry hamper on Sandra's desk with a ribbon around it and a note that says: "Love, your lab mates".

Background text: Alice is a graduate student who works in a lab with other graduate students. One of Alice's colleagues, Sandra has a very frustrating habit of hoarding all of the lab's sticky notes in her office. The sticky notes are meant for all of the lab members and everyone often needs them to carry out their work. Alice and the other lab mates have pointed out the situation to Sandra but they don't know how to exactly say the right thing and Sandra keeps hoarding all of the lab's sticky notes in her office. Alice and the other lab mates decide to do something about this.

A.1.6 Sweaty clothes

	Costly	*Costless*
Informative	The next day before Sandra arrives at the lab, Alice leaves a pile of sweaty and stinky clothes on Sandra's desk with a ribbon around them and a note that says: "Love, your lab mates".	The next day before Sandra arrives at the lab, Alice leaves a brand-new laundry hamper on Sandra's desk with a ribbon around it and a note that says: "Love, your lab mates".
Uninformative	The next day before Sandra arrives at the lab, Alice covers Sandra's entire desk in sticky notes and leaves a note on the desk that says: "Love, your lab mates".	The next day before Sandra arrives at the lab, Alice leaves a giant cardboard box full of sticky notes with a ribbon around them on Sandra's desk along with a note that says: "Love, your lab mates".

Background text: Alice is a graduate student who works in a lab with other graduate students. One of Alice's colleagues, Sandra, goes to the gym every morning and has a very frustrating habit of drying her sweaty, stinky towel on the lab radiator. Alice and all of her lab mates have noticed that the lab often smells badly because of this. Alice and her lab mates have pointed out the situation to Sandra but they don't know how to exactly say the right thing and Sandra keeps drying her towel on the lab radiator, stinking up the lab space. Alice and the other lab mates decide to do something about this.

A.1.7 Loose hair

	Costly	*Costless*
Informative	The next day before Steve gets back home from work, Andrew vacuums all of Steve's hair and leaves the contents of the vacuum on Steve's bed along with a note that says: "Love, your roommates".	The next day before Steve gets back home from work, Andrew leaves a brand-new hairnet on Steve's bed with a ribbon around it and a note that says: "Love, your roommates".
Uninformative	The next day before Steve gets back home from work, Andrew takes out Steve's freshly washed wet clothes out of the washer. Andrew then leaves the clothes on the dirty floor of Steve's bedroom along with a note that says: "Love, your roommates".	The next day before Steve gets back home from work, Andrew leaves a voucher for a new laundry detergent on Steve's bed with a ribbon on it along with a note that says: "Love, your roommates".

Background text: Andrew lives in a house with four other roommates. One of them, Steve, has long hair and has a very frustrating habit of throwing his loose, broken hair around the house. As a result, the entire apartment is often covered with Steve's hair. The roommates have pointed out the hair to Steve but they don't know how to exactly say the right thing and he keeps throwing his hair all around the place. Andrew talks with his other roommates and together they decide to do something about it.

A.1.8 Laundry detergent

	Costly	*Costless*
Informative	The next day before Steve gets back home from work, Andrew takes out Steve's freshly washed wet clothes out of the washer. Andrew then leaves the clothes on the dirty floor of Steve's bedroom along with a note that says: "Love, your roommates".	The next day before Steve gets back home from work, Andrew leaves a voucher for a new laundry detergent on Steve's bed with a ribbon on it along with a note that says: "Love, your roommates".
Uninformative	The next day before Steve gets back home from work, Andrew vacuums all of Steve's hair and leaves the contents of the vacuum on Steve's bed along with a note that says: "Love, your roommates".	The next day before Steve gets back home from work, Andrew leaves a brand-new hairnet on Steve's bed with a ribbon around it and a note that says: "Love, your roommates".

Background text: Andrew lives in a house with four other roommates. One of them, Steve, has a very frustrating habit of using other roommates' laundry detergent. There is a washer in the unit and all the roommates use their own laundry detergent, except Steve who uses everybody else's. The roommates have pointed this out to Steve but they don't know how to exactly say the right thing and he keeps using their laundry detergent every week. Andrew talks with his other roommates and together they decide to do something about it.

A.2 Vignettes used in Experiment 3

A.2.1 Sponge Scenario

Sandra lives in an apartment with one other roommate, Alice. Sandra found Alice on an apartment share app and the two of them have been living together for the last few months. They don't hang out much together, but they are cordial to one another. Sandra is an engineer at a big software company. She's been doing that job for over a year and seems to enjoy it. Alice recently switched jobs and started working for a company that produces kitchen cleaning supplies. She joined as a marketing manager and is responsible for creating advertising campaigns for their products. Both Sandra and Alice have demanding jobs. So, at the start of their living arrangement, they decided to divide the house chores between themselves to ensure that the apartment stays nice and clean. Alice's chores are to take the trash out and vacuum the apartment once a week. Sandra's chores are to do the dishes and clean the kitchen, which she typically has to do once every 2-3 days (to prevent the mess from piling up). One day, Sandra comes back home to find a brand-new kitchen sponge and a bottle of dish soap on her pillow with a ribbon around them and a tag that says" "Love, your roommate".

A.2.2 Dirty dishes scenario

Sandra lives in an apartment with one other roommate, Alice. Sandra found Alice on an apartment share app and the two of them have been living together for the last few months. They don't hang out much together, but they are cordial to one another. Sandra is an engineer at a big software company. She's been doing that job for over a year and seems to enjoy it. Alice recently switched jobs and started working for a company that produces kitchen cleaning supplies. She joined as a marketing manager and is responsible for creating advertising campaigns for their products. Both Sandra and Alice have demanding jobs. So, at the start of their living

arrangement, they decided to divide the house chores between themselves to ensure that the apartment stays nice and clean. Alice's chores are to take the trash out and vacuum the apartment once a week. Sandra's chores are to do the dishes and clean the kitchen, which she typically has to do once every 2-3 days (to prevent the mess from piling up).One day, Sandra comes back home to find a stack of dirty dishes on her pillow with a ribbon around them and a tag that says" "Love, your roommate".

A.2.3 Deodorant scenario

Sandra is a graduate student. She shares her office with another graduate student, Alice. Sandra and Alice have been office mates for a few months now. They get along well but keep mostly to themselves. Sandra and Alice have very different personalities. Sandra's desk is covered with stacks of papers, and piles of books. On the other hand, Alice's desk is clear and organized. Her books are stacked neatly in the shelf above her desk and her papers are stored neatly in her files. The only things on Alice's desk are her computer, a potted plant, and a box of beauty products from a subscription service that delivers a new box to her every week. The box contains all sorts of products ranging from personal hygiene (soaps, deodorants, etc.) to beauty (lipsticks, eyeshadows, etc.). Sandra and Alice also have different working schedules. Sandra goes to the gym every morning. She takes her shower and gets ready there after her workout. Sandra comes to work straight from the gym. As a result, she always has her gym clothes and gym towel with her and leaves them to dry on the office radiator, every day. Alice, on the other hand, comes to work first thing in the morning. She is always dressed impeccably and always has great perfume on her. Alice and Sandra both mind their own business and keep to themselves, but they are cordial with one another, as they have to share a small space with each other. One

day, Sandra comes into work to find a brand-new bottle of deodorant on her desk, with a ribbon around it and a tag that says: "Love, your officemate".

A.2.4 Sweaty clothes scenario

Sandra is a graduate student. She shares her office with another graduate student, Alice. Sandra and Alice have been office mates for a few months now. They get along well but keep mostly to themselves. Sandra and Alice have very different personalities. Sandra's desk is covered with stacks of papers, and piles of books. On the other hand, Alice's desk is clear and organized. Her books are stacked neatly in the shelf above her desk and her papers are stored neatly in her files. The only things on Alice's desk are her computer, a potted plant, and a box of beauty products from a subscription service that delivers a new box to her every week. The box contains all sorts of products ranging from personal hygiene (soaps, deodorants, etc.) to beauty (lipsticks, eyeshadows, etc.). Sandra and Alice also have different working schedules. Sandra goes to the gym every morning. She takes her shower and gets ready there after her workout. Sandra comes to work straight from the gym. As a result, she always has her gym clothes and gym towel with her and leaves them to dry on the office radiator, every day. Alice, on the other hand, comes to work first thing in the morning. She is always dressed impeccably and always has great perfume on her. Alice and Sandra both mind their own business and keep to themselves, but they are cordial with one another, as they have to share a small space with each other. One day, Sandra comes into work to find a pile of sweaty and stinky clothes on her desk, with a ribbon around them and a tag that says: "Love, your officemate".

A.2.5 Headphones scenario

Scott is a software engineer. He works in an office and sits on a desk which is surrounded by only one other desk. That one desk belongs to one of Scott's colleagues, Adam. Scott and

Adam get along well, but usually keep to themselves and their own work. They work in different divisions of the company and have very different social interests. Scott loves football and in his free time at the office, he loves to talk to others about recent games. Adam, on the other hand, loves video games. He has a blog where he discusses new games and various gaming equipment, like controllers, wireless headsets, and headphones. Adam's blog is so successful that when different gaming companies come out with new gaming equipment, they often send the products to Adam for free, so that he can review them and write about them on his blog. Adam and Scott also have very different working styles. Adam loves to get into work early and get a head start on the day, whereas Scott comes in late and stays late to finish his work. Adam likes to work in complete peace and quiet, but Scott loves to listen to music loudly on his portable speakers during work time. One day, Scott gets into work to find a brand-new set of headphones on his desk with a ribbon around them and a tag that says: "Love, your desk mate".

A.2.6 Batteries scenario

Scott is a software engineer. He works in an office and sits on a desk which is surrounded by only one other desk. That one desk belongs to one of Scott's colleagues, Adam. Scott and Adam get along well, but usually keep to themselves and their own work. They work in different divisions of the company and have very different social interests. Scott loves football and in his free time at the office, he loves to talk to others about recent games. Adam, on the other hand, loves video games. He has a blog where he discusses new games and various gaming equipment, like controllers, wireless headsets, and headphones. Adam's blog is so successful that when different gaming companies come out with new gaming equipment, they often send the products to Adam for free, so that he can review them and write about them on his blog. Adam and Scott also have very different working styles. Adam loves to get into work early and get a head start on

the day, whereas Scott comes in late and stays late to finish his work. Adam likes to work in complete peace and quiet, but Scott loves to listen to music loudly on his portable speakers during work time. One day, Scott gets into work to find that the batteries in his portable speakers have been removed and in their place is a note that says: "Love, your desk mate".

A.2.7 Shampoo bottles scenario

Sandra lives in an apartment with one other roommate, Alice. Sandra found Alice on an apartment share app and the two of them have been living together for the last few months. They don't hang out much together, but they are cordial to one another. Sandra is an engineer at a big software company. She's been doing that job for over a year and seems to enjoy it. Alice recently switched jobs and started working for a company that produces hair care supplies (like shampoos and conditioners). She joined as a marketing manager and is responsible for creating advertising campaigns for their products. Sandra and Alice have very different personalities. Sandra is laid back whereas Alice is strict. Sandra is easy going with her things. She likes to share her stuff and use other people's stuff. On the other hand, Alice hates sharing. She does not like to use other people's products and does not like to share her things with others one bit, especially products for personal hygiene. As a result, at the start of their living arrangement, Alice and Sandra decided that each of them will use only their own products (like body washes, creams, shampoos) for personal care. One day, Sandra comes back home to find a brand bottle of a shampoo on her pillow with a ribbon around it and a tag that says" "Love, your roommate".

A.2.8 Empty bottles scenario

Sandra lives in an apartment with one other roommate, Alice. Sandra found Alice on an apartment share app and the two of them have been living together for the last few months. They don't hang out much together, but they are cordial to one another. Sandra is an engineer at a big

software company. She's been doing that job for over a year and seems to enjoy it. Alice recently switched jobs and started working for a company that produces hair care supplies (like shampoos and conditioners). She joined as a marketing manager and is responsible for creating advertising campaigns for their products. Sandra and Alice have very different personalities. Sandra is laid back whereas Alice is strict. Sandra is easy going with her things. She likes to share her stuff and use other people's stuff. On the other hand, Alice hates sharing. She does not like to use other people's products and does not like to share her things with others one bit, especially products for personal hygiene. As a result, at the start of their living arrangement, Alice and Sandra decided that each of them will use only their own products (like body washes, creams, shampoos) for personal care. One day, Sandra comes back home to find a basket filled with lots of empty shampoo bottles on her pillow with a ribbon around them and a tag that says "Love, your roommate".

A.3 Vignettes used in Experiment 4

A.3.1. Groceries scenario

<table>
<tr><td colspan="2">Background Information

Alice, Becky, and Cassandra work together at an office. They all work in the same division and have office desks close to one another. Alice and Becky both bring home cooked lunches to work. They both store their food in the common office fridge, along with all of the other employees.

Cassandra does not bring any lunch to work. Initially, she used to buy her lunch from the office cafeteria. However, for some time now, she has been eating Becky's lunch. Everyday just before lunch time, Cassandra takes out Becky's lunch from the common fridge and eats it, without asking Becky for her permission. This has been going on for quite some time and Cassandra has eaten many of Becky's lunches, without buying her any lunch in return or paying her back in any way.

At first Becky didn't know who was behind her missing lunch but recently, Becky found out that the culprit was Cassandra. She has tried to speak to Cassandra about it. But Cassandra continues to eat Becky's lunch without asking her and without paying her for the food.

Alice is aware of the whole situation and decides to do something about it.</td></tr>
<tr><td>Benefits: second party (victim)

Alice knows Cassandra credit card details. Alice uses Cassandra's credit card to buy a month's worth of groceries for Becky. She gets all of the groceries delivered to Becky at her home and leaves the receipt on Cassandra's desk.</td><td>Benefits: third party (bystander)

Alice knows Cassandra credit card details. Alice uses Cassandra's credit card to buy a month's worth of groceries for herself. She gets all of the groceries delivered to herself at her home and leaves the receipt on Cassandra's desk.</td></tr>
</table>

A.3.2. Uber scenario

<table>
<tr><td colspan="2">Background Information

Alice, Becky, and Cassandra live together in an apartment. They started living together a couple years ago when they all moved to the city to start new jobs. All three of them have offices that are close to their apartment. However, Cassandra, who is an architect, has to travel far into the city, twice a week for a project she’s working on.

On the days that Cassandra has to travel far for work, she takes an Uber. However, Cassandra does not have the Uber app on her phone. Instead, when she needs a ride, she books her Uber off of Becky's app. Becky is the only one in the apartment with the app. However, Cassandra books the rides without asking Becky. In fact, not only does she use Becky's app, she also uses Becky's credit card to book the rides for herself. This has been going on for some time and Cassandra has taken rides costing close to $100 without paying Becky back.

Becky recently found out about this Uber situation. She has tried to speak to Cassandra about it. But Cassandra continues to use Becky's app without asking her and without paying her money back.

Alice is aware of the whole situation and decides to do something about it.</td></tr>
<tr><td>Benefits: second party (victim)

Alice knows that Cassandra has an Amazon Prime account. In fact, Alice has both the login details and the credit card details for Cassandra's account. Alice uses Cassandra's Amazon account and Cassandra’s credit card to buy Becky a new $100 jacket that Becky has been wanting to get. She gets the jacket delivered to Becky at home and leaves the receipt on Cassandra's bed.</td><td>Benefits: third party (bystander)

Alice knows that Cassandra has an Amazon Prime account. In fact, Alice has both the login details and the credit card details for Cassandra's account. Alice uses Cassandra's Amazon account and Cassandra's credit card to buy herself a new $100 jacket that she has been wanting to get. She gets the jacket delivered to herself at home and leaves the receipt on Cassandra's bed.</td></tr>
</table>

A.3.3. Clothes scenario

<table>
<tr><td colspan="2">Background Information

Alice, Becky, and Cassandra live together in an apartment. They started living together a couple years ago when they all moved to the city to start new jobs. All three of them have their own rooms and closets. They share only the essentials in the kitchen and otherwise, use only their own things.

However, Cassandra has been taking Becky's expensive, formal clothes out of her closet and wearing them. Alice's clothes are too big for her, but Becky's clothes fit Cassandra perfectly. So, Cassandra has been taking Becky's expensive clothes, without asking Becky. In fact, not only has she been taking Becky's clothes, she has also ruined some of them, by leaving stains and tears in them. This has been going on for sometime and Cassandra has taken many formal clothes from Becky without replacing them or paying Becky back.

Becky recently found out about this clothes situation. She has tried to speak to Cassandra about it. But Cassandra continues to take Becky's clothes without asking her and without paying her for the damages.

Alice is aware of the whole situation and decides to do something about it.</td></tr>
<tr><td>Benefits: second party (victim)

Alice knows Cassandra credit card details. Alice uses Cassandra's credit card to buy Becky a brand-new iPhone, since Becky's phone had become old and needed to be replaced. She gets the phone delivered to Becky at home and leaves the receipt on Cassandra's bed.</td><td>Benefits: third party (bystander)

Alice knows Cassandra credit card details. Alice uses Cassandra's credit card to buy herself a brand-new iPhone, that she has been wanting to get. She gets the phone delivered to herself at home and leaves the receipt on Cassandra's bed.</td></tr>
</table>

A.3.4. Stationary scenario

<table>
<tr><td colspan="2">Background Information

Alice, Becky, and Cassandra work together at an office. They all work in the same division and have office desks close to one another. Alice and Becky both use their own stationeries and work supplies, like most other employees.

Cassandra, however, does not bring her own stationery. Instead, she has been using Becky's stationery, without asking Becky for permission. Every morning, Cassandra gets in early and takes whatever stationery she fancies from Becky's desk and then claims it as her own. This has been going on for quite some time and Cassandra has taken many of Becky's notepads, sticky notes, highlighters, and pens without paying her back in any way.

At first Becky didn't know who was behind her missing stationery but recently, Becky found out that the culprit was Cassandra. She has tried to speak to Cassandra about it. But Cassandra continues to take Becky's stationery without asking her and without paying her for the supplies.

Alice is aware of the whole situation and decides to do something about it.</td></tr>
<tr><td>Benefits: second party (victim)

Alice knows Cassandra credit card details. Alice uses Cassandra's credit card to buy a yearlong supply of stationery for Becky. She gets all of the stationery delivered to Becky at the office and leaves the receipt on Cassandra's desk.</td><td>Benefits: third party (bystander)

Alice knows Cassandra credit card details. Alice uses Cassandra's credit card to buy a yearlong supply of stationery for herself. She gets all of the stationery delivered to herself at the office and leaves the receipt on Cassandra's desk.</td></tr>
</table>

A.4 Experiment 5

Method

Participants Participants were recruited using Amazon's Mechanical Turk. 2010 people completed the Experiment. Of these, 1682 passed our attention checks (16% excluded; 52% female, modal age range was 31-40 years). Participants were paid $0.33.

Design & materials We modified the text of the scenarios used in Experiment 1 to put participants in the place of the punisher. For instance, for the dirty dishes vignette, Sandra is described as the participant's roommate (e.g., *your roommate, Sandra*) and the punishment is described as being enacted by the participant (e.g., *you decide to do something about it*; *you leave a brand-new dish sponge on her pillow*, etc.). As in Experiment 1, participants read only one of the 32 vignettes and answered the same four questions (using the same rating scales) to assess their expectations of the effectiveness of their punishment. Each scenario also included a comprehension question and participants failing to answer this question correctly were excluded from all analyses.

Results

The first question asked participants how likely it was that Sandra would get the message. (Figure A.4.1(a)). Like Experiment 1, we found a main effect of cost imposed, $b = 0.50$, $t_{1671.56} = 3.48$, $p < .001$, a main effect of informativeness, $b = 3.24$, $t_{1671.77} = 22.46$, $p < .001$, but unlike Experiment 1, we did not find a significant cost × informativeness interaction, $b = -0.38$, $t_{1671.07} = -1.31$, $p = .19$. However, pairwise contrasts revealed results qualitatively similar to those of Experiment 1. When the punishment was informative and communicative, the cost it imposed did not matter, as participants rated the recipient to be equally likely to get the message from costly, literal punishment (e.g., dirty dishes; $M = 6.14$, $SE = 0.23$, CI: [5.64, 6.64]) as from a

costless, figurative punishment (e.g., new sponge and dish-soap; $M = 5.83$, $SE = 0.23$, CI: [5.33, 6.32]), $t(1671) = -1.53$, $p = .13$. However, when the punishment was uninformative, participants judged the recipient to be more likely to get the message from a costly punishment (e.g., mail thrown around messily; $M = 3.09$, $SE = 0.23$, CI: [2.60, 3.59]), compared with a costless punishment (e.g., mail stacked away neatly; $M = 2.40$, $SE = 0.23$, CI: [1.91, 2.90]), $t(1671) = -3.40$, $p < .001$.

Next, we looked at the perceived effectiveness of each punishment in bringing about a change in the recipient's future behavior (Figure A.4.1(b)). The results revealed a main effect of cost, $b = 0.41$, $t_{1671.92} = 2.95$, $p = .003$ and of informativeness, $b = 2.23$, $t_{1672.25} = 15.89$, $p < .001$. There was no significant interaction between the two, $b = -0.30$, $t_{1671.17} = -1.08$, $p = .28$. Pairwise contrasts showed that when punishments were informative, costless punishments ($M = 4.94$, $SE = 0.19$, CI: [4.54, 5.35]) were perceived to be as effective in changing behavior as costly punishment ($M = 5.20$, SE = 0.19, CI: [4.80, 5.61]), $t(1671) = -1.32$, $p = .19$. However, when punishments were uninformative, the cost imposed made a difference as costly punishments ($M = 3.13$, $SE = 0.19$, CI: [2.72, 3.53]) were perceived to be more effective in changing behavior than costless punishments ($M = 2.56$, $SE = 0.19$, CI: [2.16, 2.96]), $t(1672) = -2.86$, $p = .004$.

Participants' judgments regarding the recipient's feelings upon receiving the punishments also showed a main effect of cost imposed, $b = -2.14$, $t_{1672.72} = -17.03$, $p < .001$, and a significant interaction, $b = 1.20$, $t_{1671.32} = 4.78$, $p < .001$, however, no significant effect of informativeness, $b = -0.20$, $t_{1673.30} = -1.56$, $p = .12$ (Figure A.4.1(c)). Pairwise contrasts revealed that, when the punishment imposed a cost, participants expected the recipient to feel worse when the punishment was informative ($M = 3.03$, $SE = 0.15$, CI: [2.73, 3.32]) than when it was uninformative ($M = 2.62$, $SE = 0.15$, CI: [2.33, 2.92]), $t(1942) = -2.27$, $p = .02$. Similar results

were found for the costless condition, such that participants judged the recipient to feel significantly worse when the punishment was informative and directly related to the transgression ($M = 4.56$, $SE = 0.15$, CI: [4.27, 4.86]) than when it was uninformative and unrelated to the transgression ($M = 5.36$, $SE = 0.15$, CI: [5.07, 5.65]), $t(1672) = 4.51$, $p < .001$.

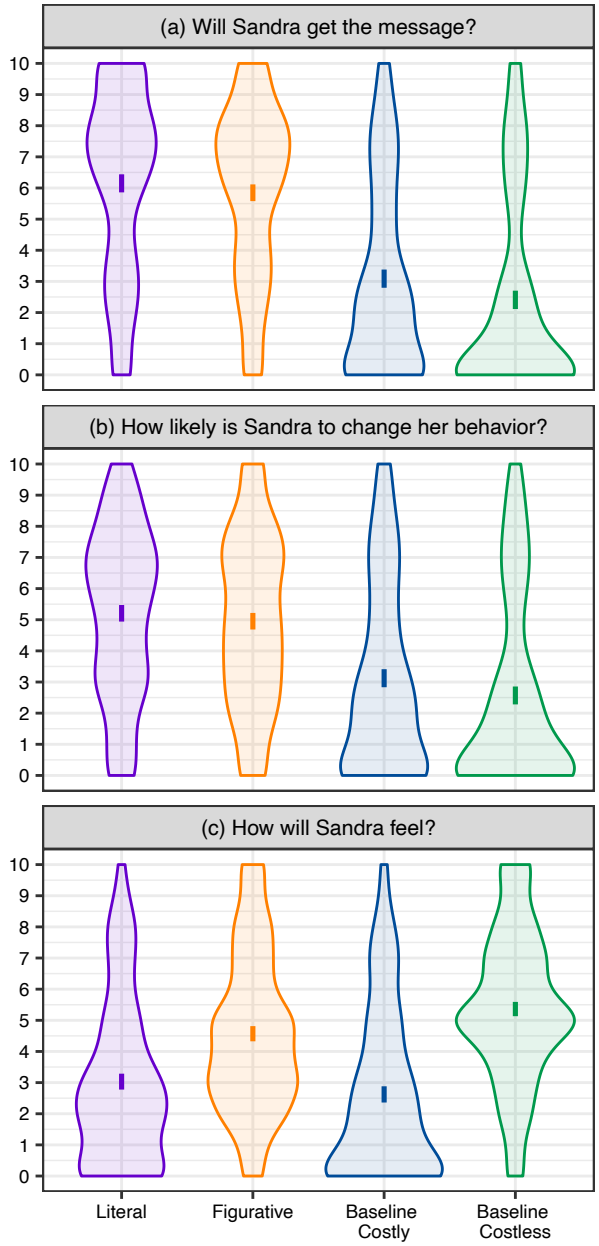

Figure A.4.1: Distribution of participant responses for three of the four dependent variables in Experiment 1. Lines inside the violins represent 95% CIs of means.

Finally, we examined ratings of whether the act in question would feel like a punishment (results are on the log odds ratio scale). We found a main effect of cost imposed, $b = 1.23$, $SE = 0.11$, $z = 11.65$, $p < .001$, informativeness, $b = 0.72$, $SE = 0.11$, $z = 6.90$, $p < .001$, and a significant interaction, $b = -0.85$, $SE = 0.21$, $z = -4.03$, $p < .001$ (see Figure A.4.2). Pairwise contrast revealed that, when the punishment was costly, informative punishments ($M = 0.77$, $SE = 0.15$, CI: [0.49, 1.07]) were judged to feel more like punishments that uninformative punishments ($M = 0.48$, $SE = 0.15$, CI: [0.19, 0.76]), $z = -2.06$, $p = .04$. A similar pattern of results was found for costless punishments, as informative punishments were judged to feel more like punishments ($M = -0.03$, $SE = 0.14$, CI: [-0.31 0.25]) than uninformative punishments ($M = -1.18$, $SE = 0.15$, CI: [-1.49, -0.88]), $z = -7.63$, $p < .001$.

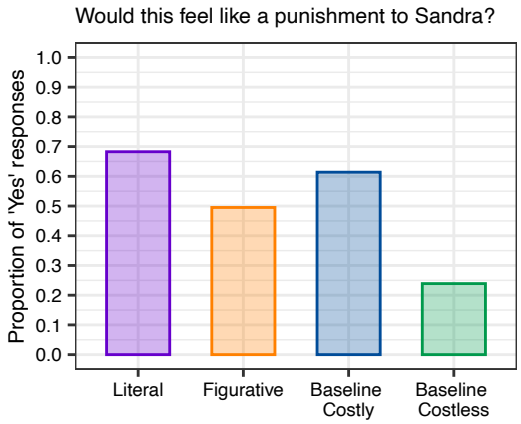

Figure A.4.2: Proportion of participants who think the action would feel like a punishment to the target.

A.5 Experiment 6

Method

Participants 2023 participants completed the study. From these 1518 participants passed the attention check (25% excluded; 58% women; modal age range between 31-40 years). Participants were recruited using Amazon's Mechanical Turk and paid $0.33.

Design & materials We used the same set of scenarios as in Experiment 1 and Experiment 5. The only difference was in the perspective employed. All scenarios were described by putting the participant in the transgressor's perspective. Participants read, for instance, that they had become lax at doing their dishes and came back home one day to find a brand-new kitchen sponge (or a stack of their dirty dishes) on their bed. The item left varied on the same two dimensions as before – it was either informative and related to the transgression or uninformative and unrelated to it, and it either imposed a cost or was costless. Like before, participants read only one of the 32 vignettes and answered the same four questions (using the same rating scales) as in Experiments 1 and 5 to assess their expectations of the effectiveness of their punishment. Each scenario also included a comprehension question and participants failing to answer this question correctly were excluded from all analyses.

Results

Like Experiment 1, we found a main effect of cost imposed, $b = 1.00$, $t_{1507.93} = 6.15$, $p < .001$, a main effect of informativeness, $b = 3.31$, $t_{1508.41} = 21.41$, $p < .001$, and a significant interaction between the two $b = -1.44$, $t_{1507.85} = -4.67$, $p < .001$ for ratings of how likely participants (now in the transgressor's perspective) would be to get the message. Pairwise contrasts revealed results similar to those of Experiments 1 and 5. When the punishment was informative and communicative, the cost it imposed did not matter, as participants rated

themselves to be equally likely to get the message from costly, literal punishment (e.g., dirty dishes; $M = 8.56$, $SE = 0.18$, CI: [8.18, 8.93]) as from a costless, figurative punishment (e.g., new sponge and dish-soap; $M = 8.33$, $SE = 0.18$, CI: [7.96, 8.70]), $t(1508) = -1.04$, $p = .30$. However, when the punishment was uninformative, participants reported being more likely to get the message from a costly punishment (e.g., mail thrown around messily; $M = 5.97$, $SE = 0.18$, CI: [5.60, 6.34]), compared with a costless punishment (e.g., mail stacked away neatly; $M = 4.30$, $SE = 0.19$, CI: [3.92, 5.67]), $t(1508) = -7.68$, $p < .001$.

Participant's judgments about their likelihood of changing their future behavior exhibited a similar pattern of results (Figure A.5.1(b)). We found a main effect of cost, $b = 0.88$, $t_{1507.68} = 5.95$, $p < .001$, of informativeness, $b = 2.49$, $t_{1508.06} = 16.91$, $p < .001$ and a significant interaction between the two, $b = -1.23$, $t_{1507.61} = --4.19$, $p < .001$. Pairwise contrasts showed that when punishments were informative, costless punishments ($M = 8.17$, $SE = 0.19$, CI: [7.79, 8.55]) were perceived to be as effective in changing behavior as costly punishments ($M = 8.43$, $SE = 0.19$, CI: [8.05, 8.81]), $t(1508) = -1.24$, $p = .22$. However, when punishments were uninformative, the cost imposed made a difference as costly punishments ($M = 6.56$, $SE = 0.18$, CI: [6.18, 6.93]) were perceived to be more effective in changing behavior than costless punishments ($M = 5.06$, $SE = 0.19$, CI: [4.68, 5.45]), $t(1507) = -7.20$, $p < .001$.

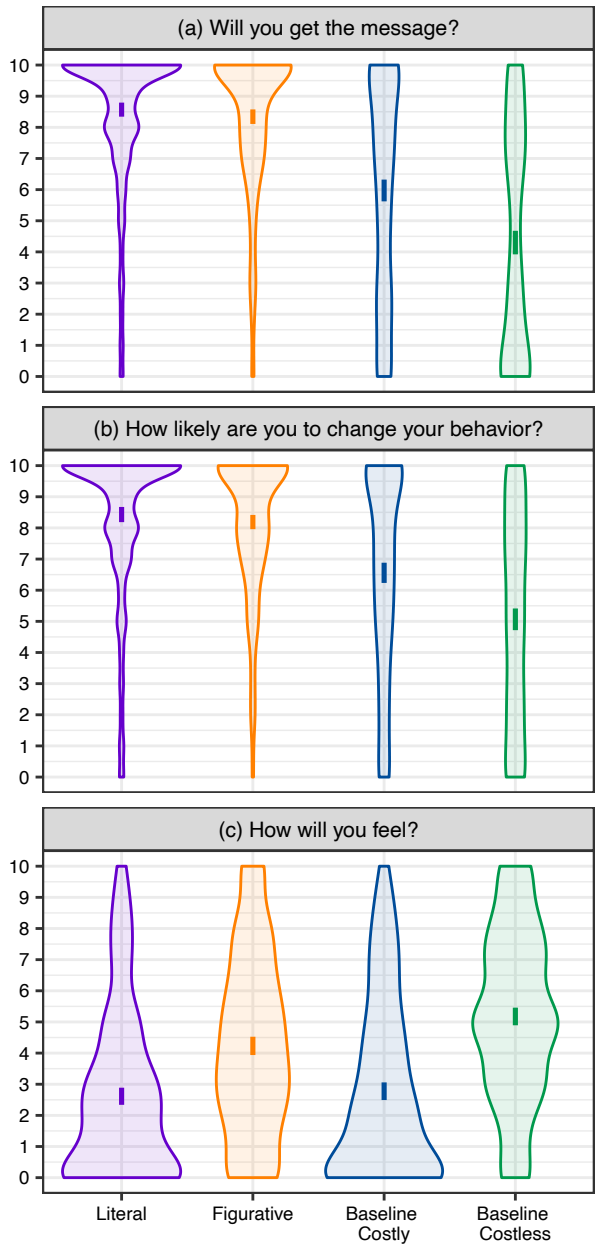

Figure A.5.1: Distribution of participant responses for three of the four dependent variables in Experiment 1. Lines inside the violins represent 95% CIs of means.

Results for participant's judgments about how they would feel on receiving the items also revealed a main effect of cost imposed, b = -2.02, $t_{1507.81}$= -14.06, p < .001, a main effect of informativeness, b = -0.55, $t_{1508.15}$= -3.84, p < .001, as well as a significant interaction between cost and informativeness, b = 0.77, $t_{1507.75}$ = 2.68, p = .008. Pairwise contrasts revealed that, when the punishment imposed a cost, participants judged themselves to feel bad irrespective of

whether the punishment was informative (M = 2.61, SE = 0.19, CI: [2.22, 3.00]) or uninformative (M =2.77, SE = 0.18, CI: [2.39, 3.15]), $t(1508) = 0.83$, $p = .41$. However, when the punishment was costless, they judged themselves to feel significantly worse when it was informative and directly related to the transgression (M = 4.24, SE = 0.19, CI: [3.86, 4.62]) than when it was uninformative and unrelated to the transgression (M =5.18, SE = 0.19, CI: [4.97, 5.65]), $t(1508) = 4.57$, $p < .001$.

Finally, we examined ratings of whether the act in question would feel like a punishment (results are on the log odds ratio scale) to the participants. We found a main effect of cost imposed, $b = 1.27$, $SE = 0.11$, $z = 11.53$, $p < .001$ and of informativeness, $b = 0.30$, $SE = 0.11$, $z = 2.67$, $p = .008$, but no significant interaction, $b = -0.33$, $SE = 0.22$, $z = -1.52$, $p = .13$ (Figure A.5.2). Pairwise contrast revealed that, when the punishment was costly, both informative punishments (M = 0.83, SE = 0.15, CI: [0.53, 1.13]) and uninformative punishments (M = 0.70, SE = 0.15, CI: [0.41, 1.00]), were judged to feel like punishments, $z = -0.80$, $p = .42$. However, when punishments were costless, informative punishments (M = -0.28, SE = 0.15, CI: [-0.56 0.02]) were judged to feel more like punishments than uninformative punishments (M = -0.73, SE = 0.15, CI: [-1.03, -0.43]), $z = -3.00$, $p = .003$.

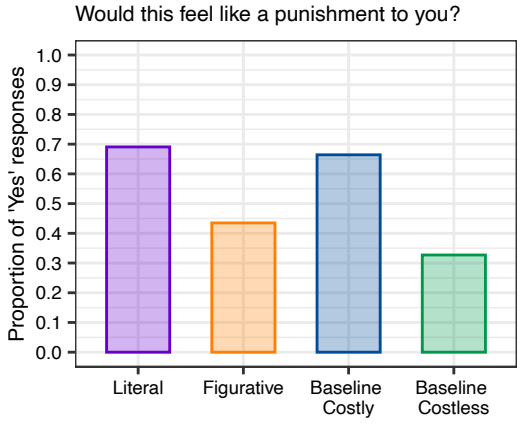

Figure A.5.2: Proportion of participants who think the action would feel like a punishment to the target.

A.6 Bayesian hierarchical model for Experiment 1 data

In the main text of the manuscript, we employ a model that has random intercepts only for the stimuli. As is recommended by some (Barr et al., 2013), we attempted to model our data to include both random intercepts and slopes for the items (i.e., models with 'maximal' random effect structure). These maximally specified models failed to converge, following which we estimated each of our models through Bayesian hierarchical modeling with uninformative priors. Using brms (Bürkner, 2017), we ran models that included random intercepts and slopes for items, along with fixed effects for cost, informativeness, and their interaction (DV~ cost*informativeness + (cost*informativeness|item), data = data).

We first looked at participants' expectations about each punishment and how clearly it would allow the recipient to 'get the message'. Results revealed that the Bayesian 95% credible interval did not overlap with 0 for both main effects of cost [0.06, 0.94] and informativeness [2.94, 4.61] but did overlap with 0 for their interaction [-2.38, 0.38].

Next, we looked at participant ratings of how likely the target was to modify their behavior following the punishment. 95% credible intervals did not overlap with 0 for the cost [0.15, 0.95], for informativeness [2.01, 3.05]; but did overlap with 0 for the interaction between cost and informativeness [-1.83, 0.47].

How the target would feel, after witnessing the punishment, was the next question and we found that the Bayesian 95% credible interval did not overlap with 0 for cost [-3.01, -1.91] but did overlap with 0 for the interaction of cost with informativeness [-0.38, 2.35] and for informativeness [-1.59, 0.78]. However, when we looked at each scenario individually, informativeness was significant for all but one of them.

Finally, we looked at whether participants would judge the action to feel like a punishment to the target. Bayesian 95% credible intervals did not overlap with 0 for cost [1.25, 1.89], or for informativeness [0.13, 1.75] but did overlap with 0 for their interaction [-1.39, 0.16].

Results presented here converge qualitatively with those presented in the main text, with some minor discrepancies. These discrepancies, we believe are a result of overparameterization. Bates et al., (Bates et al., 2015) caution that convergence failures (like the ones we encountered) signal overfitting and even when converge is reached (for instance, by using Bayesian hierarchical modeling), overparametrization may lead to poor estimation procedures and produce models that are uninterpretable. It is perhaps also worth noting that a major difference between our simpler models and these Bayesian Hierarchical models was regarding the 95% credible interval for the interaction term which overlapped with 0 for each of the four dependent measures. In the simpler models reported in the main text, the interaction term revealed that when the punishment was informative the cost it imposed did not matter. However, when it was uninformative, costly punishments were judged to be more effective than costless punishments. Results from the Bayesian Hierarchical Models thus change our interpretation of the effectiveness of the costly, uninformative punishment, suggesting that these are no more effective than costless uninformative punishments. These results do not however alter our main finding that when a punishment is communicative, its cost does not matter as costless informative punishments are judged to be as effective in communicating the message and getting the perpetrator to change their behavior as costly informative punishments.

Correlation Tables from Experiment 1

APPENDIX A

Table A.6.1

Literal Punishment

Measure	1	2	3	4
1. Message	---			
2. Future behavior	0.64	---		
3. Feel	-0.15	-0.12	---	
4. Punish	0.31	0.23	-0.40	---

Table A.6.2

Figurative Punishment

Measure	1	2	3	4
1. Message	---			
2. Future behavior	0.71	---		
3. Feel	-0.30	-0.14	---	
4. Punish	0.27	0.18	-0.50	---

Table A.6.3

Baseline Costly Punishment

Measure	1	2	3	4
1. Message	---			
2. Future behavior	0.79	---		
3. Feel	0.15	0.11	---	
4. Punish	0.17	0.16	-0.34	---

Table A.6.4

Baseline Costless Punishment

Measure	1	2	3	4
1. Message	---			
2. Future behavior	0.82	---		
3. Feel	-0.16	-0.14	---	
4. Punish	0.32	0.29	-0.53	---

A.7 Experiment 7

We tested whether the nature of the punisher-transgressor relationship affected the punishment a punisher chose. We explored two relational dynamics: (a) friends vs strangers, and (b) landlord vs tenant.

Methods

Participants 165 participants took part in Experiment 7a. Of these, 160 participants (3% excluded; 54% female, modal age range 31-40 years) completed the study and passed the comprehension check. Experiment 7b also had an initial sample of 165, of which 153 (7% excluded; 60% female, modal age range 25-30 years) participants completed the study and passed the comprehension check. All participants were recruited through Amazon Mechanical Turk and were paid $0.20 in return for their participation.

Design & materials For both a & b we modified the vignette used in Experiment 2b. The basic set-up was the same: participants were asked to imagine that their roommate, Sandra, has not been doing the dishes and as the punisher they have to find a way to respond to her transgression by choosing one of following three options: talk to Sandra, punish her figuratively, or punish her literally. The main difference was in the relationship they shared with Sandra. In Experiment 7a participants were told that Sandra was either one of their closest friends or that they had come to know and live with her recently. In Experiment 7b they were told that Sandra was either the owner of the apartment and their landlord or that they were the owner of the apartment and Sandra was their roommate and tenant. The experiment was fully between-subjects and participants were assigned to read only one condition.

Results

For the friends vs strangers condition, a chi-square test of goodness-of-fit found evidence in favor of the null hypothesis: proportion of choice was similar across the two between-subjects condition $X^2(2) = 2.01$, $p = .37$, $V_{Cramer} = 0.00$, *CI*: [-0.20, 0.04], $n = 160$ (Figure A.7.1). When Sandra was described as a friend, most participants preferred to talk to her (73%, *CI*: [64%, 83%]), followed by punishing her figuratively (25%, *CI*: [16%, 35%]), and literally (2%, *CI*: [0%, 13%]). When Sandra was described as a stranger, the pattern of choice was similar: 76% preferred to talk (*CI*: [67%, 85%]), 24% preferred to punish figuratively (*CI*: [15%, 33%]), and no one wanted to punish literally (*CI*: [0%, 9%]).

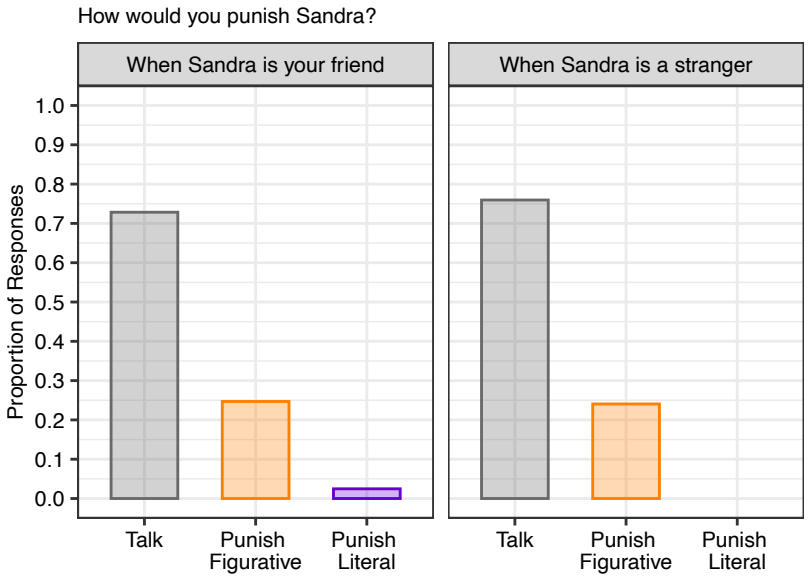

Figure A.7.1: Proportions reflecting which punishment participants would choose themselves as punishers, broken down by the two experimental conditions.

For the landlord vs tenant condition, a chi-square test of goodness-of-fit also found evidence in favor of the null hypothesis: X^2 (2) = 0.71, $p = .70$, $V_{Cramer} = 0.00$, *CI*: [-0.18, 0.08], $n = 153$ (Figure A.7.2). When confronting Sandra, the landlord, 78% (*CI*: [70%, 87%]) chose to talk to her, 21% (*CI*: [13%, 31%]) chose to use figurative punishment, and 1% (*CI*: [0%, 11%]) chose the literal punishment. When confronting Sandra, the tenant, about 81% (*CI*: [73%, 89%])

opted to talk, followed by 17% (*CI*: [9%, 26%]) choosing to punish figuratively followed by 3% (*CI*: [0%, 11%]) punishing literally.

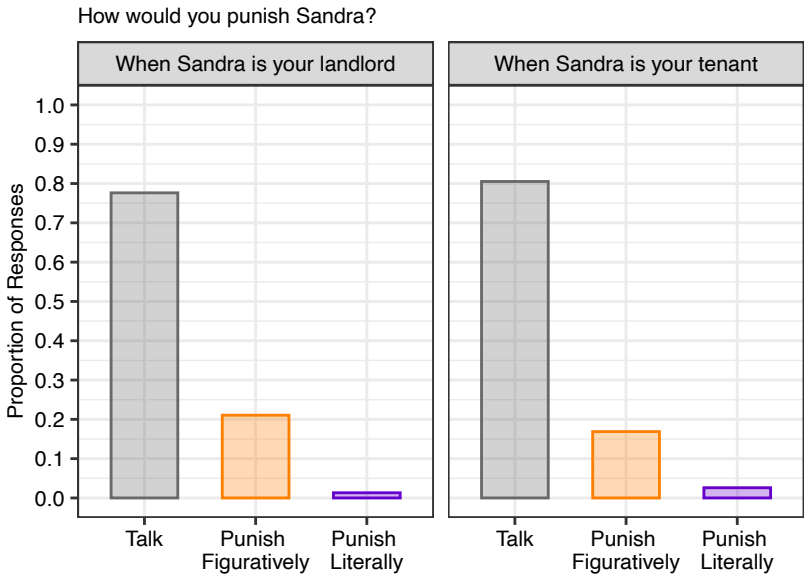

Figure A.7.2: Proportions reflecting which punishment participants would choose themselves as punishers, broken down by the two experimental conditions.

A.8 Experiment 8

The aim of this experiment was to test an alternative explanation for people's preference for figurative punishments reported in Experiment 2a. When given a choice between literal, figurative, baseline costly, and baseline costless punishments, people displayed an overwhelming preference to punish the perpetrator figuratively. One explanation for this pattern of results is that people prefer to punish in ways that enhance the communicative signal of their punishment while imposing minimal costs. An alternative explanation for the result is that the choice has nothing to do with the communicative value of the punishment and instead is a reflection of the practicality of figurative punishments. Each instance of figurative punishment bestows upon the recipient an item they can use to address the problem at hand. For instance, giving your roommate a brand-new sponge and dish-soap gets them one step closer to doing the dishes. It could be then that the reason people prefer figurative punishments is for their practicality rather than their communicative value. In this experiment we put this hypothesis to test.

Methods

Participants 101 participants took part in the Experiment. Of these, 92 participants (9% excluded; 55% female, modal age range 31-40 years) completed the study and passed the comprehension check. All participants were recruited through Amazon Mechanical Turk and were paid $0.25 in return for their participation.

Design & materials We used the same set of vignettes as Experiment 2a. Participants read one of the eight vignettes and were asked how they would respond to the norm-violating agent. They were given the choice to punish the perpetrator literally, figuratively, using the baseline costly punishment, or the baseline costless punishment. The main difference was however that all items used for figurative punishments were now 'toy' items. For instance, in the dirty dishes scenario

instead of leaving Sandra a brand-new kitchen sponge and dish soap, participants were given the option to leave her a toy, plastic sponge and dish soap. This change was made to take away the practical utility of the figurative punishments. The toy items retained their semantic connection to the violation at hand and carried a clear communicative inference but lacked any practical value. If participants' preference for figurative punishments stemmed from their practicality, then, the pattern of results from this experiment should differ from those of Experiment 2a. If instead, participants' preference is tied to the communicative value of figurative punishment then, the pattern of results from this experiment should echo those of Experiment 2a.

Results

Collapsing across different contexts, a chi-square test of goodness-of-fit was performed to determine whether the four kinds of punishments were preferred equally. We found evidence in favor of the alternative hypothesis X^2 (3) = 121.30, $p < .001$, $V_{Cramer} = 0.66$, *CI*: [0.56, 0.76], $n = 92$. Descriptive analysis revealed that overall, figurative punishments were selected most frequently (73%, *CI*: [62%, 83%]), followed by literal punishments (22%, *CI*: [14%, 31%]), baseline costless punishments (4%, *CI*: [0%, 14%]) and baseline costly punishments (2%, *CI*: [0%, 11%]) (Figure A.8.1). The preference for figurative punishments held in every context (except for one, see Table A.8.1). Chi-square tests conducted simultaneously within each scenario, using the ggstatsplot package (Patil, 2021), provided evidence for the alternative hypothesis and a preference for the figurative punishment for all but one context (see Table A.8.1). The results from this experiment mirror the results of Experiment 2a, demonstrating that participants' preference for figurative punishments is tied to their communicative value.

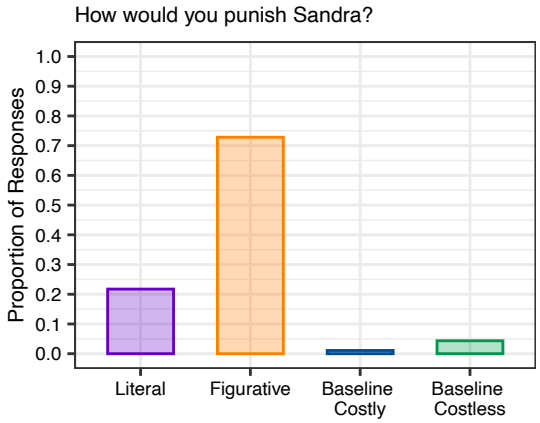

Figure A.8.1: Proportions reflecting which punishment participants would choose themselves as punishers.

Table A.8.1
Choice decision broken down by scenario.

		Percent selected						
Context	N	Literal	Figurative	Baseline Costly	Baseline Costless	Chi-square statistic	p.value	df
Sweaty clothes	11	9.09%	72.72%	9.09%	9.09%	13.4	3.91e-3	3
Stationary	10	40%	60%	0	11.11%	10.8	1.29e-2	3
Loud music	12	25%	66.67%	0	8.33%	12.7	5.42e-3	3
Messy mail	11	45.45%	45.45%	0	9.09%	7.55	5.64e-2	3
Laundry	12	8.33%	91.67%	0	0	28.7	2.63e-6	3
Hair	12	16.67%	83.33%	0	0	22.7	4.74e-5	3
Fridge food	13	15.38%	84.62%	0	0	25.5	1.24e-5	3
Dirty dishes	11	18.18%	72.72%	0	9.09%	14.1	2.78e-3	3

B

Supplementary materials for Chapter 2

B.1 Vignettes used in Experiment 1

B.1.1 Context: Food

(*Common background*): Katherine is an employee at a company. The head of the company asks Katherine to plan an end-of-year dinner to celebrate their accomplishments. To do so, Katherine works with a local restaurant to prepare for this event. The restaurant shows Katherine a menu for her approval. Katherine takes a look at the menu. She sees that one of the dishes is a pasta dish with some anchovies blended into the tomato sauce for a salty flavor. As it happens, one of the employees, Jan, is quite allergic to anchovies.

(*Accidental*) Katherine does not know this since Jan never told her. She approves of the pasta on the menu. (*Negligent*) Katherine knows this because Jan had once told her. However, when Katherine is looking over the food menu, she forgets about Jan's allergy. She approves of the pasta on the menu. (*Intentional*) Katherine knows this because once Jan had told her. When Katherine is looking over the food menu, she remembers Jan's allergy to the anchovies. Nonetheless, she approves of the pasta on the menu.

(*Harmful outcome*): On the night of the party, the final menu includes the anchovy-based pasta dish to which Jan is allergic. As soon as Jan takes one bite of the pasta, she has a severe allergic reaction. Jan is rushed to the hospital and ends up spending the entire night under critical care. (*Harmless outcome*): On the night of the party, the final menu includes the anchovy-based pasta dish to which Jan is allergic. Jan has a last-minute change to her plans, however, and ends up giving the company party a miss.

B.1.2 Context: Cleaning

(*Common background*): Adam is a professional cleaner and has recently joined a new cleaning company. His company cleans spaces professionally, but they only use cleaning

supplies that they make themselves. Soon after he joins, Adam is sent over to a client's house to mop and shine their floors. He is given a bag of the company's self-made supplies and is asked to use only those to do the cleaning. When Adam reaches the client's house, he notices that some of the floors are made with pure hardwood. Adam knows that pure hardwood floors need to be cleaned only with a gentle, chemical-free, cleaner. He carefully examines his supply bag and picks out the cleaner labelled 'gentle' to do the cleaning. It turns out that the cleaning company's self-made gentle cleaner contains a small but significant percentage of chemical bleach in it (a fact not mentioned on the bottle).

(*Accidental*) Adam does not know that the company's homegrown gentle cleaner has bleach in it because no one told Adam and it is not mentioned anywhere on the bottle. He chooses the bleach infused gentle cleaner to do the cleaning. He chooses the bleach infused gentle cleaner to do the cleaning. (*Negligent*) Adam knows that the gentle cleaner has bleach in it because his boss had told him. However, as Adam is selecting a cleaner, he fails to recall this fact regarding the bleach. He chooses the bleach infused gentle cleaner to do the cleaning. He chooses the bleach infused gentle cleaner to do the cleaning. (*Intentional*) Adam knows that the gentle cleaner has bleach in it because his boss had told him. As Adam is selecting a cleaner, he recalls this fact regarding the bleach. Nonetheless, he chooses the bleach infused gentle cleaner to do the cleaning.

(*Harmful outcome*): The bleach in the cleaner reacts harshly with the hardwood floors stripping the pigment and varnish off the wood. The floor is ruined. (*Harmless outcome*): As Adam picks up the cleaner to start, he realizes that bottle in the supply bag given to him is damaged and almost empty. He decides to use regular gentle soap and warm water to clean the floors. This does the trick and leaves the floors looking shiny and new.

B.1.3 Context: Laptop

(*Common background*): Simon is a personal assistant to the CEO of a large multinational corporation. Simon's boss is giving an important presentation later in the day, but his charger has stopped working. He asks Simon to find him a laptop charger. Simon takes a look at the laptop and the charging port. This particular laptop can only handle up-to 60-watts of charge at once.

(*Accidental*) Simon does not know this because he is unfamiliar with this laptop and his Boss never mentioned the charging restrictions. Simon picks a common 100-watt charger and decides to use that to charge the laptop. (*Negligent*) However, at the time of picking a charger, Simon forgets about the laptop's charging capacity and picks a 100-watt charger and decides to use that on the laptop. (*Intentional*) Simon knows this because his Boss had mentioned this to him. Simon realizes that he needs to plug only a 60-watt charger into the laptop. Yet, he picks a 100-watt charge and decides to use that on the laptop.

(*Harmful outcome*): Simon plugs in the 100-watt charger on the 60-watt laptop. The power surge is too much. The laptop heats up and breaks down. (*Harmless outcome*): As Simon opens his Boss's laptop to plug the 100-watt charger, he notices that it is fully charged. His Boss also decides to make some last-minute changes and ends up doing the presentation without having to charge the laptop at all.

B.1.4 Context: Aquarium

(*Common background*): Mary is a primary school teacher, teaching 6-year-olds. Every year, the school takes their 6-years-olds for a field trip to the aquarium at the outskirts of the town. This year, Mary is helping out with the field trip. The aquarium is very strict about regulating the number of people inside at any given point in time. For this reason, tickets are booked in slots

and no one is allowed to enter without a ticket. The schools' headteacher purchased tickets for all the 6-year-olds months in advance.

(*Accidental*) A Day before the trip, the headteacher hands Mary a large pile of documents. She however forgets to tell Mary that the tickets for the trip are in one of the envelopes in the pile. On the morning of field trip, as Mary is getting ready to leave, she does not know that she needs to the tickets put the envelope with all of the tickets in her bag. She leaves the house without the tickets. (*Negligent*) She lets Mary know that the tickets for the trip are in one of the envelopes in the pile. On the morning of the field trip, as Mary is getting ready to leave, she forgets that she needs to put the envelope with all of the tickets in her bag. She leaves the house without the tickets. (*Intentional*) She lets Mary know that the tickets for the trip are in one of the envelopes in the pile. On the morning of the field trip, as Mary is getting ready to leave, she remembers that she needs to put the envelope with all of the tickets in her bag. Nonetheless she decides to leave the house without the tickets.

(*Harmful outcome*): Mary and the rest of the teachers and students get to the aquarium at the designated time. Mary does not have the tickets for entry. As a result, the class is not allowed to enter the aquarium and the trip is ruined. (*Harmless outcome*): Mary and the rest of the teachers and students get to the aquarium at the designated time. Mary does not have the tickets for entry. However, the aquarium had a plumbing incident in the morning and had to shut down temporarily for the day. As a result, the aquarium ends up rescheduling the school's visit.

B.1.5 Context: Cake

(*Common background*): Zoey's sister Molly is pregnant and is having a baby shower at her place. The shower is in intimate gathering of Molly's closest friends and family and is a potluck. Molly has asked Zoey to bring the cake. Zoey goes to the grocery store to get all the ingredients

for the cake. When she reaches the aisle with the icing, Zoey reaches out for the tub of chocolate icing. It turns out that Molly cannot have chocolate icing. Since the start of the pregnancy, anything with cocoa and chocolate makes Molly severely nauseous.

(*Accidental*) Zoey does not know this because Molly never told her. Zoey picks the tub of chocolate icing and decorates the cake with it. (*Negligent*) Zoey knows this well because Molly had told her. However, at the time of picking the icing, Zoey forgets about Molly's intolerance and nausea. She picks the tub of chocolate icing and decorates the cake with it. (*Intentional*) Zoey knows this well because Molly had told her. Yet, Zoey decides to pick the tub of chocolate icing and decorates the cake with it.

(*Harmful outcome*): On the day of the baby shower, Zoey reaches the party with her chocolate iced cake. The smell and sight of the cake induce severe nausea in Molly and Molly spends her entire baby shower throwing up in the bathroom. (*Harmless outcome*): On the day of the baby shower, Zoey reaches the party with her chocolate iced cake. Molly however ends up having to make a last-minute trip to the hospital due to pre-mature labor pains. The shower ends up getting postponed and Zoey goes back home with her chocolate cake.

B.1.6 Context: Gardening

(*Common background*): Tim is a gardener. He was hired by Anna. Anna has a garden in her backyard that she wants to use to grow organic fruits and vegetables. She has been trying to grow the plants herself but has so far been unsuccessful. She has hired Tim to help her. After speaking to Anna over the phone and accepting the job, Tim goes over to Anna's place. He inspects the garden and realizes that the plants need more fertilizers to grow. He carefully examines the plants and the supplies he has in his bag. He picks a new fertilizer he has bought recently. The

fertilizer is mostly organic but does have some strong industrial chemicals. It turns out that Anna wants Tim to use only organic manures and fertilizers in the garden.

(*Accidental*) Tim does not know about Anna's desire to use only organic supplies because Anna never told him. Tim picks the new fertilizer that is mostly organic but has some strong industrial chemicals in it. (*Negligent*) Tim knows this because Anna told him at the time of hiring him. As Tim is deciding on a fertilizer, he forgets Anna's request. He picks the new fertilizer that is mostly organic but has some strong industrial chemicals. (*Intentional*) Tim knows this because Anna told him at the time of hiring him. As Tim is deciding on a fertilizer, he remembers Anna's request. Yet, he decides to use the new fertilizer than is mostly organic but has some strong industrial chemicals in it.

(*Harmful outcome*): Tim uses the new fertilizer on the plants. The chemicals are too strong for the plants, and they die. The soil bed is also ruined. (*Harmless outcome*): When Tim opens the bottle to use the fertilizer, he realizes that he doesn't have the mixing liquid required. He ends up using a different, older fertilizer that happens to be 100% organic.

B.2 Vignettes used in Experiment 2

B.2.1 Context: Food

Time 1: Negligent Harm

(*Common background*): Katherine is an employee at a company. The head of the company asks Katherine to plan an end-of-year dinner to celebrate their accomplishments. To do so, Katherine works with a local restaurant to prepare for this event. The restaurant shows Katherine a menu for her approval. Katherine takes a look at the menu. She sees that one of the dishes is a pasta dish with some anchovies blended into the tomato sauce for a salty flavor. As it happens, one of the employees, Jan, is quite allergic to anchovies. Katherine knows this because Jan had

once told her. However, when Katherine is looking over the food menu, she forgets about Jan's allergy. She approves of the pasta on the menu. On the night of the party, the final menu includes the anchovy-based pasta dish to which Jan is allergic. As soon as Jan takes one bite of the pasta, she has a severe allergic reaction. Jan is rushed to the hospital and ends up spending the entire night under critical care.

Time 2:

(Common background): Six months have passed, and Katherine has been asked to plan another company party. She works with a local restaurant to curate a menu for the event. One of the dishes on the menu is a broccoli salad made with an anchovy-dressing. One of the employees in the company, Jan, is quite allergic to anchovies. Katherine knows this well.

(Negligent): Despite that, at the time of finalizing the menu Katherine forgets about Jan's allergy again. Katherine has no intention of hurting Jan and would have removed the salad from the menu had she remembered Jan's allergy. However, at the time of finalizing the menu, Katherine completely forgets about Jan's allergy and approves the anchovy-based salad. On the night of the party, the anchovy-based salad is on the menu. *(Harm)* Jan takes a bite of the salad and immediately has an allergic reaction. She is rushed to the hospital again and ends up spending the entire night there. *(No Harm)* Jan takes a bite of the salad but has no allergic reaction. Unknown to anyone, Jan had taken a precautionary anti-allergy pill before coming to the party. She has the salad and is unharmed.

(Not Negligent): Owing to that, at the time of finalizing the menu Katherine recalls Jan allergy. Katherine has no intention of hurting Jan and recalls Jan's allergy when she sees the anchovy dressing. At the time of finalizing the menu, Katherine remembers Jan's allergy and removes the anchovy-based salad. On the night of the party, a different (non-anchovy-based

dressing) is part of the menu. *(Harm)* However, as soon as Jan takes a bite of the salad, she has an allergic reaction. Unknown to anyone, Jan is also allergic to lemon poppy seeds that the salad contained. She is rushed to the hospital where she ends up spending the night. (*No Harm)* Jan takes a bite of the salad and has no allergic reaction. She enjoys all the food, including the salad and has a good time at the company party.

B.2.2 Context: Cleaning

Time 1: Negligent Harm

(*Common background*): Adam is a professional cleaner and has recently joined a new cleaning company. His company cleans spaces professionally, but they only use cleaning supplies that they make themselves. Soon after he joins, Adam is sent over to a client's house to mop and shine their floors. He is given a bag of the company's self-made supplies and is asked to use only those to do the cleaning. When Adam reaches the client's house, he notices that some of the floors are made with pure hardwood. Adam knows that pure hardwood floors need to be cleaned only with a gentle, chemical-free, cleaner. He carefully examines his supply bag and picks out the cleaner labelled 'gentle' to do the cleaning. It turns out that the cleaning company's self-made gentle cleaner contains a small but significant percentage of chemical bleach in it (a fact not mentioned on the bottle). Adam knows that the gentle cleaner has bleach in it because his boss had told him. However, as Adam is selecting a cleaner, he fails to recall this fact regarding the bleach. He chooses the bleach infused gentle cleaner to do the cleaning. The bleach in the cleaner reacts harshly with the hardwood floors stripping the pigment and varnish off the wood. The floor is ruined.

Time 2:

(Common background): Six months have passed, and Adam is sent to another client's house to clean their sinks. When he arrives, he learns that the sinks in the house are somewhat old, and notes that the drains will be easily damaged by any chemicals. Adam picks out the cleaner from his bag labeled 'gentle'.

(Negligent): Adam knows that this cleaner has bleach in it because his boss has told him. Adam has no intention of damaging the client's drains. Despite that, at the time of cleaning the sinks, Adam again forgets that the gentle cleaner contains harsh chemicals. He chooses the chemical infused cleaner. Adam cleans the sinks using the cleaner containing bleach. *(Harm)* When the client goes to turn on the sink in their kitchen, water starts leaking from the drain. The harsh chemicals in the cleaner damaged the drains in the house. The client is very unhappy with Adam's work and the drains are ruined. *(No Harm)* But when the client goes to inspect their sinks, the drains work perfectly. Unknown to anyone, Adam's boss had recently changed their cleaner mixtures, and so the chemicals that Adam used did not damage the drains in the house. The client is very happy with Adam's work, and they thank him for doing an excellent job.

(Not Negligent): But before he begins to clean the sinks, Adam remembers that the gentle cleaner contains a small percentage of chemical bleach, which would damage the drains in the house. Adam has no intention of damaging the client's drains, so he puts the cleaner back into his bag and chooses a cleaner he knows contains no harsh chemicals instead. Adam cleans the sinks in the house using a safe, chemical free cleaner. *(Harm)* But when the client goes to turn on the sink in their kitchen, water starts leaking from the drain. Unknown to anyone, the client's neighbors who share the same drainpipe, disposed of a large amount of food waste in their kitchen sink. The greasiness of the food damages the drains, and the drains are ruined. (*No*

Harm) When the client sees their newly cleaned sinks, they are very happy with Adam's work, and they thank him for doing a great job.

B.2.3 Context: Laptop

Time 1: Negligent Harm

(*Common background*): Simon is a personal assistant to the CEO of a large multinational corporation. Simon's boss is giving an important presentation later in the day, but his charger has stopped working. He asks Simon to find him a laptop charger. Simon takes a look at the laptop and the charging port. This particular laptop can only handle up-to 60-watts of charge at once. Simon knows this because his Boss had mentioned this to him. However, at the time of picking a charger, Simon forgets about the laptop's charging capacity and picks a 100-watt charger and decides to use that on the laptop. Simon plugs in the 100-watt charger on the 60-watt laptop. The power surge is too much. The laptop heats up and breaks down

Time 2:

(Common background): Six months have passed, and Simon's boss has an important conference call scheduled for later in the day. He needs his computer to attend the call as well as to reference the notes for the meeting. He asks Simon to make sure his laptop is charged for the meeting. Simon knows the laptop can only handle up to 60-watts of charge at once, because his boss has mentioned this to him before. Simon has no intention of damaging his boss's computer.

(Negligent): However, when he goes to plug in the laptop, he forgets about the laptop's charging limit again and attaches a 100-watt charger. *(Harm)* After Simon plugs the 100-watt charger into the laptop, the laptop overheats and breaks down. As a result, Simon's boss cannot look up his notes and cannot attend the conference call. *(No Harm)* Simon plugs the 100-watt charger into the laptop, and the laptop charges normally. Unknown to Simon, his boss had

recently gotten his computer battery replaced with a new battery that can handle a greater amount of charge. Simon's boss is able to attend the conference call and he thanks Simon for charging his laptop.

(Not Negligent): He grabs the 100-watt charger, but when he goes to plug in the laptop, he remembers the laptop's charging limit. *(Harm)* Simon instead plugs a 60-watt charger into the laptop. Suddenly, the computer screen goes blank, and the computer dies. Unknown to Simon, his boss had recently gotten his hard drive replaced, and the replacement drive caused the computer to freeze up when connected to certain chargers. As a result, Simon's boss cannot attend the conference call. (*No Harm)* Simon instead plugs a 60-watt charger into the laptop, and the laptop charges normally. As a result, Simon's boss is able to attend the conference call and he thanks Simon for charging his laptop.

B.2.4 Context: Aquarium

Time 1: Negligent Harm

(*Common background*): Mary is a primary schoolteacher, teaching 6-year-olds. Every year, the school takes their 6-years-olds for a field trip to the aquarium at the outskirts of the town. This year, Mary is helping out with the field trip. The aquarium is very strict about regulating the number of people inside at any given point in time. For this reason, tickets are booked in slots and no one is allowed to enter without a ticket. The schools' headteacher purchased tickets for all the 6-year-olds months in advance. A day before the trip, the headteacher hands Mary a large pile of documents. She lets Mary know that the tickets for the trip are in one of the envelopes in the pile. On the morning of the field trip, as Mary is getting ready to leave, she forgets that she needs to put the envelope with all of the tickets in her bag. She leaves the house without the tickets. Mary and the rest of the teachers and students get to the aquarium at the designated time.

Mary does not have the tickets for entry. As a result, the class is not allowed to enter the aquarium and the annual trip to the aquarium is ruined.

Time 2:

(Common background): Six months have passed and Mary has been asked to help out with the school's field trip once again. The day before the trip, the headteacher places an envelope containing all the 6-year-old students' tickets in Mary's filing cabinet and asks Mary to bring the tickets along the next day. Mary knows the aquarium is very strict about requiring tickets, and she makes note of the fact that the envelope with the tickets is in the filing cabinet. She has no intention of disrupting the school's field trip.

(Negligent): However, the next morning, as she grabs her things from the classroom before the field trip, she forgets about the tickets again. Mary leaves without her students' tickets to the aquarium. *(Harm)* When the students arrive at the aquarium, they aren't allowed to enter. The field trip is cancelled, and the visit to the aquarium is ruined. *(No Harm)* However, when the students arrive at the aquarium, they discover that the aquarium is offering free admission that day. Although they don't have their tickets, the students are all able to enter the aquarium.

(Not Negligent): The next morning, as she grabs her things from the classroom before the field trip, she remembers the envelope of tickets and grabs it from the filing cabinet. Mary puts the envelope containing her students' tickets in her bag. *(Harm)* However, when the students arrive, they discover that the aquarium is closed. It turns out that many of the employees were down with the flu. So, owners of the aquarium had temporarily shut it down but had not notified the school. The field trip is cancelled, and the visit to the aquarium is ruined. (*No Harm)* When the students arrive at the aquarium, they show their tickets and are all able to enter. The field trip goes well.

B.3 Vignettes used in Experiment 3

B.3.1 Context: Food

(Common background): Katherine is an employee at a company. The head of the company asks Katherine to plan an end-of-year dinner to celebrate their accomplishments. To do so, Katherine works with a local restaurant to prepare for this event. The restaurant shows Katherine a menu for her approval.

(Precaution): Two weeks prior to the event, Katherine sends an email to all the employees in the company to ask about their allergies and dietary restrictions. She makes a list of all the replies and double-checks her list against the menu. *(No Precaution):* In the weeks prior to the event, Katherine does not ask any of the employees in the company about their allergies or dietary restrictions. She finalizes the menu on her own.

(Outcome): One of the dishes on the menu is a pasta dish with some anchovies blended into the tomato sauce for a salty flavor. As it happens, one of the employees, Jan, is quite allergic to anchovies. Katherine knows this because Jan had emailed it to her (or because Jan had told her). However, as she is looking over the food menu, Katherine fails to recall that Jan has an allergy to anchovies. As a result, she approves of the pasta dish. However, as she is looking over the food menu, Katherine fails to recall that Jan has an allergy to anchovies. As a result, she approves of the pasta dish. The final menu includes the pasta dish made with anchovies to which Jan is allergic. However, on the night of the party, the restaurant has to substitute the pasta dish with a salad due to issues with their food stock. As a result, the pasta dish with the anchovies is not part of the dinner. Jan also has a last-minute change to her plans and she ends up giving the company party a miss.

B.3.2 Context: Cleaning

(*Common background*): Adam is a professional cleaner and has recently joined a new cleaning company. His company cleans spaces professionally, but they only use cleaning supplies that they make themselves. Soon after he joins, Adam is sent to mop and shine a client's floors. He is given a bag of the company's self-made supplies and is asked to use only those to do the cleaning. Since he is unfamiliar with the products, Adam sits down with all of the company's cleaning supplies and makes himself a chart that lists all of the ingredients for each product.

(Precaution): Since he is unfamiliar with the products, Adam sits down with all of the company's cleaning supplies and makes himself a chart that lists all of the ingredients for each product. When Adam reaches the client's house, he sees that some of the floors are made with pure hardwood. Adam knows that pure hardwood floors must be cleaned only with a gentle, chemical-free cleaner. He checks the chart that he made when he was hired and decides to use a cleaner labelled 'gentle' to do the cleaning. *(No Precaution):* Adam is unfamiliar with products but decides to play it by ear and figure out the supplies when he gets to the client's house.

(Outcome): It turns out that the cleaning company's self-made gentle cleaner contains a small but significant percentage of chemical bleach in it (a fact not mentioned on the bottle). Adam knows that the gentle cleaner has bleach in it because his boss had mentioned this to him. However, when he takes the bottle from his bag, he forgets that the 'gentle' cleaner contains harsh chemicals, and he decides to use it to clean the floors. Adam is about to begin cleaning the floors with the unsafe cleaner when Adam's client suddenly remembers that they have an important event to attend that day and asks Adam to reschedule. As a result, a different employee returns to the client's house later that week and properly cleans the floors.

B.3.3 Context: Laptop

(*Common background*): Simon is a personal assistant to the CEO of a large multinational corporation. Simon's boss is giving an important presentation later in the day but his charger has stopped working. He asks Simon to find him a laptop charger.

(Precaution): Before searching for a charger, Simon examines the laptop's charging port and takes a picture of it to be sure he finds a compatible charger. He checks the specification on the bottom of the laptop and finds that the laptop is compatible with a 100-watt charger. He goes to the storeroom and finds a 100-watt laptop charger for his boss's computer. *(No Precaution):* Simon glances at the laptop and then goes to the storeroom to find whichever laptop charger is lying around to plug in to his boss' computer. He finds a 100-watt charger lying around and decides to use that for his boss's computer.

(*Outcome):* It turns out that Simon's boss had to get his laptop repaired some time back and as a result it can handle only 60 watts of charge at once. Simon knows this because his boss had mentioned it to him once before. However, at the time of selecting a charger, Simon completely forgets about the laptop's charging capacity. He picks the 100-watt charger to plug into his boss's laptop that can handle only 60 watts. As Simon opens his boss's laptop to plug in the 100-watt charger, he notices that the laptop is already fully charged. His boss ends up having to present sooner than expected, and as a result, the charger never gets plugged into the laptop.

B.3.4 Context: Aquarium

(*Common background*): Mary is a primary school teacher, teaching 6-year-olds. Every year, the school takes their 6-years-olds for a field trip to the aquarium at the outskirts of the town. It is a fun day out for the students and something they look forward to all year round. This year, Mary has agreed to help with the logistics of the field trip. The aquarium is very strict about regulating

the number of people inside at any given point in time. For this reason, tickets are booked in slots and no one is allowed to enter without a ticket. The schools' headteacher purchased tickets for all the 6-year-olds months in advance and gave them all to Mary.

(Precaution): Mary sets a reminder on her phone to make sure she will remember the tickets on the day of the field trip. On the morning of the field trip, Mary gets up early and spends time carefully going over the plan for the day and the things she will need. She sees the reminder on her phone and puts the tickets on the table so she can grab them on her way out. However, as she is leaving the house, she fails to realize that she needs to pick up the tickets for the aquarium.
(No Precaution): Mary takes the tickets home and leaves them in a pile along with her mail. On the morning of the trip, Mary wakes up late and gives little thought to the field trip that day. She rushes to leave the house. As she is getting ready, she fails to realize that she needs to pack the tickets for the aquarium.

(Outcome): She leaves the house without the tickets. Mary and the rest of the teachers and students get to the aquarium at their designated time. Mary does not have the tickets for entry. However, the aquarium had an unexpected incident in the morning and had to shut down temporarily for the day. As a result, the school ends up having to reschedule their visit to the aquarium.

B.3.5 Context: Cake

(*Common background*): Zoey's sister Molly is pregnant and is having a baby shower at her place. The shower is an intimate gathering of Molly's closest friends and family and is a potluck. Molly has asked Zoey to bring the cake.

(Precaution): Zoey asks one of Molly's friends about Molly's likes and dislikes before deciding which kind of cake to make. Zoey makes a detailed list of all the ingredients she will

need to bake a cake for Molly and goes to the grocery store to pick them up. When she reaches the aisle with the icing, Zoey reaches out for the tube of chocolate icing. *(No Precaution):* Zoey does not spend very much time thinking about which kind of cake to make – she decides to go with the first recipe she finds. Zoey then wings a trip to the grocery store and picks up items she thinks she will need whenever she happens to see them on the aisles. When she reaches the aisle with the icing, Zoey reaches out for the tube of chocolate icing.

(Outcome): It turns out that Molly cannot have chocolate icing. Since the start of the pregnancy, anything chocolate makes Molly severely nauseous. Zoey knows this well because Molly had told her. However, at the time of picking the icing, Zoey forgets about Molly's intolerance and selects the tube of chocolate icing. She makes the cake with the chocolate icing. On the day of the baby shower, Zoey makes her way to the party with her chocolate iced cake. Molly however ends up having to make a last-minute trip to the hospital due to premature labor pains. The shower ends up getting postponed and Zoey goes back home with her chocolate cake.

B.3.6 Context: Gardening

(Common background): Tim is a gardener. He was hired by Anna. Anna has a garden in her backyard that she wants to use to grow organic fruits and vegetables. She has been trying to grow the plants herself but has so far been unsuccessful. She has hired Tim to help her.

(Precaution): After speaking to Anna over the phone and accepting the job, Tim makes a note of all the work Anna wants done in her yard. He then goes over to her house and inspects the plants and the soil, making notes for himself. He tells Anna that the plants need more fertilizer and starts to look through his manual to find a fertilizer that would be good for Anna's garden. He decides to go with a new fertilizer he purchased recently. The fertilizer is mostly organic but has some chemicals in it. *(No Precaution):* After speaking to Anna over the phone

and accepting the job, Tim goes over to Anna's place to look at the garden. He briefly looks at the garden and tells Anna the plants need more fertilizer. He decides to pick any fertilizer he can find from his bag and ends up grabbing the bottle of a new fertilizer he purchased recently. The fertilizer is mostly organic but has some chemicals in it.

(Outcome): It turns out that Anna wants Tim to use only organic manures and fertilizers in the garden. Tim knows this because Anna told him this at the time of hiring him. However, as Tim is picking a fertilizer, Anna's request completely slips his mind. He picks the new fertilizer that has some chemicals in it. Tim opens the bottle to use the fertilizer only to realize that he doesn't have the mixing liquid required for the fertilizer. He ends up using a different, older fertilizer that happens to be 100% organic instead.

B.4 Vignettes used in Experiment 4

B.4.1 Context: Food

(*Common background*): Katherine is an employee at a company. The head of the company asks Katherine to plan an end-of-year dinner to celebrate their accomplishments. To do so, Katherine works with a local restaurant to prepare for this event. The restaurant shows Katherine a menu for her approval.

(High care): Katherine spends many hours deliberating each dish on the menu. She wants to make sure everyone's dietary restrictions, especially of those with food allergies, are addressed. She sees that one of the dishes is a pasta dish with some anchovies blended into the tomato sauce for a salty flavor. As it happens, one of the employees, Jan, is quite allergic to anchovies. Katherine knows this, because once Jan had told her. As she is carefully deliberating over the food menu, Katherine fails to recall that Jan has an allergy to anchovies. *(No care):* Katherine spends just a few minutes skimming over the menu. She spends little time considering the dishes

and whether everyone's dietary restrictions are addressed. She sees that one of the dishes is a pasta dish with some anchovies blended into the tomato sauce for a salty flavor. As it happens, one of the employees, Jan, is quite allergic to anchovies. Katherine knows this, because once Jan had told her. As she is quickly skimming over the food menu, Katherine fails to recall that Jan has an allergy to anchovies.

(Outcome): As a result, she approves of the pasta dish. The final menu includes the pasta dish made with anchovies to which Jan is allergic. However, on the night of the party, the restaurant has to substitute the pasta dish with a salad due to issues with their food stock. As a result, the pasta dish with the anchovies is not part of the dinner. Jan also has a last-minute change to her plans and she ends up giving the company party a miss.

B.4.2 Context: Cleaning

(*Common background*): Adam is a professional cleaner and has recently joined a new cleaning company. His company cleans spaces professionally, but they only use cleaning supplies that they make themselves. Soon after he joins, Adam is sent to mop and shine a client's floors. He is given a bag of the company's self-made supplies and is asked to use only those to do the cleaning. It turns out that the cleaning company's self-made gentle cleaner contains a small but significant percentage of chemical bleach in it (a fact not mentioned on the bottle). Adam knows that the gentle cleaner has bleach in it because his boss had told him. However, as he is examining the bottle he picked, he forgets that the 'gentle' cleaner contains any harsh chemicals, and he decides to use it to clean the floors.

(High care): When Adam reaches the client's house, he pays close attention to the floors, and observes that some of them are made with pure hardwood. Adam knows that pure hardwood floors need to be cleaned only with a gentle, chemical-free, cleaner. Adam likes his job and

wants to make sure to do a good job. He carefully examines his supply bag and picks out the cleaner labelled ‘gentle’ to do the cleaning. It turns out that the cleaning company’s self-made gentle cleaner contains a small but significant percentage of chemical bleach in it a fact not mentioned on the bottle. Adam knows that the gentle cleaner has bleach in it because his boss had told him. However, when he grabs the bottle from his bag, he forgets that the ‘gentle’ cleaner contains any harsh chemicals, and he decides to use it to clean the floors. *(No care):* When Adam reaches the client’s house, he hastily glances around and happens to notice that some of the floors are made with pure hardwood. Adam knows that pure hardwood floors need to be cleaned only with a gentle, chemical-free, cleaner. Adam does not enjoy his job. He wants to quickly finish the cleaning and go home. He shuffles through his supply bag and grabs the first cleaner he finds. The bottle is labelled ‘gentle’.

(Outcome): Adam is about to begin cleaning the floors with the unsafe cleaner when Adam’s client suddenly remembers that they have an important event to attend that day and asks Adam to reschedule. As a result, a different employee returns to the client’s house later that week and properly cleans the floors.

B.4.3 Context: Laptop

(*Common background*): Simon is a personal assistant to the CEO of a large multinational corporation. Simon’s boss is giving an important presentation later in the day but his charger has stopped working. He asks Simon to find him a laptop charger.

(High care): Simon admires his boss and likes his work. He cares about the project his boss is working on and wants the presentation to go smoothly. He examines the laptop and spends time finding a high-quality spare charger for his boss. He ends up picking out a 100-watt laptop charger that is compatible with most types of computers. *(No care):* Simon despises his boss and

dislikes his work. He does not care about the project his boss is working on and the presentation. He briefly glances over the charging port and reaches for the first charger he can find. He ends up picking out a 100-watt laptop charger.

(Outcome): It turns out that Simon's boss's laptop is an old one and can handle only 60 watts of charge at once. Simon knows this because his boss had mentioned it to him once before. However, at the time of selecting a charger, Simon completely forgets about the laptop's charging capacity. He picks the 100-watt charger to plug into his boss's old laptop that can handle only 60 watts. As Simon opens his boss's laptop to plug in the 100-watt charger, he notices that the laptop is already fully charged. His boss ends up having to present sooner than expected, and as a result, the charger never gets plugged into the laptop.

B.4.4 Context: Aquarium

(*Common background*): Mary is a primary school teacher, teaching 6-year-olds. Every year, the school takes their 6-years-olds for a field trip to the aquarium at the outskirts of the town. It is a fun day out for the students and something they look forward to all year round. This year, Mary has agreed to help with the logistics of the field trip. The aquarium is very strict about regulating the number of people inside at any given point in time. For this reason, tickets are booked in slots and no one is allowed to enter without a ticket. The schools' headteacher purchased tickets for all the 6-year-olds months in advance and gave them all to Mary.

(High care): Mary is looking forward to the trip, and she wants to do everything she can to make sure the event goes smoothly. On the morning of the field trip, Mary gets up early and spends time carefully going over the plan for the day and the things she will need. As she is getting ready, she fails to realize that she needs to pack the tickets for the aquarium. She forgets about the tickets and leaves for the aquarium without them. *(No care):* Mary is not particularly

excited about the aquarium. On the morning of the trip, Mary wakes up late and gives little thought to the field trip that day. As she is getting ready, she fails to realize that she needs to pack the tickets for the aquarium. She forgets about the tickets and leaves for the aquarium without them.

(Outcome): Mary and the rest of the teachers and students get to the aquarium at their designated time. Mary does not have the tickets for entry. However, the aquarium had an unexpected incident in the morning and had to shut down temporarily for the day. As a result, the school ends up having to reschedule their visit to the aquarium.

B.4.5 Context: Cake

(*Common background*): Zoey's sister Molly is pregnant and is having a baby shower at her place. The shower is in intimate gathering of Molly's closest friends and family and is a potluck. Molly has asked Zoey to bring the cake.

(High care): Zoey needs to go to the grocery store to get all the ingredients. Before making that trip, Zoey carefully plans her trip to the store, including making a detailed list of everything she will need to make the perfect cake for Molly. When she reaches the aisle with the icing, Zoey reaches out for the tube of chocolate icing. *(No care):* Zoey does not care much about the party and spends little time thinking about which kind of cake to make. She goes to the grocery store to get all the ingredients she thinks she will need for the cake. When she reaches the aisle with the icing, Zoey reaches out for the tube of chocolate icing.

(Outcome): It turns out that Molly cannot have chocolate icing. Since the start of the pregnancy, anything with chocolate makes Molly severely nauseous. Zoey knows this well because Molly had told her. However, at the time of picking the icing, Zoey forgets about Molly's intolerance and selects the tube of chocolate icing. She makes the cake with the

chocolate icing. On the day of the baby shower, Zoey makes her way to the party with her chocolate iced cake. Molly however ends up having to make a last-minute trip to the hospital due to pre-mature labor pains. The shower ends up getting postponed and Zoey goes back home with her chocolate cake.

B.4.6 Context: Gardening

(*Common background*): Tim is a gardener. He was hired by Anna. Anna has a garden in her backyard that she wants to use to grow organic fruits and vegetables. She has been trying to grow the plants herself but has so far been unsuccessful. She has hired Tim to help her. After speaking to Anna over the phone and accepting the job, Tim goes over to Anna's place.

(High care): Tim really likes his work and enjoys making clients happy. Tim inspects the garden carefully and notes that the plants will need more fertilizers to grow. He examines the plants and the supplies in his bag to select the best fertilizer for the job. He picks a new fertilizer he has bought recently. The fertilizer is mostly organic but has some strong industrial chemicals.
(No care): Tim dislikes his work. He doesn't care much for the job and wants to do the bare minimum. He glances at the garden and tells Anna the plants need more fertilizer. He opens his bag and grabs the first bottle of fertilizer he can find. It's a new fertilizer he has bought recently. The fertilizer is mostly organic but has some strong industrial chemicals.

(Outcome): It turns out that Anna wants Tim to use only organic manures and fertilizers in the garden. Tim knows this because Anna told him this at the time of hiring him. However, as Tim grabs the fertilizer from his bag, Anna's request slips his mind. He picks the new fertilizer that has strong industrial chemicals in it. He opens the bottle to use the fertilizer only to realize that he doesn't have the mixing liquid required for the fertilizer. He ends up using a different, older fertilizer that happens to be 100% organic instead.

B.5 Vignettes used in Experiment 5

B.5.1 Context: Food

(*Common background*): Katherine is an employee at a company. The head of the company asks Katherine to plan an end-of-year dinner to celebrate their accomplishments. To do so, Katherine works with a local restaurant to prepare for this event. The restaurant shows Katherine a menu for her approval. Katherine spends many hours deliberating each dish on the menu. She wants to make sure everyone's dietary restrictions are addressed. She sees that one of the dishes is a pasta dish with some anchovies blended into the tomato sauce for a salty flavor. As it happens, one of the employees, Jan, is quite allergic to anchovies. Katherine knows this, because once Jan had told her.

(Thought: occurrent): As she is deliberating over the food menu, Katherine remembers that Jan has an allergy to anchovies. As a result, she calls the restaurant and cancels the pasta dish. On the night of the party, the final menu does not include the pasta dish that is made with anchovies to which Jan is allergic. *(Thought: not occurrent):* However, as she is deliberating over the food menu, Katherine fails to recall that Jan has an allergy to anchovies. As a result, she approves the pasta dish. The final menu includes the pasta dish made with anchovies to which Jan is allergic. However, on the night of the party, the restaurant has to substitute the pasta dish with a salad due to issues with their food stock. As a result, the pasta dish with the anchovies is not part of the dinner.

(Outcome): Jan also has a last-minute change to her plans and she ends up giving the company party a miss.

B.5.2 Context: Cleaning

(*Common background*): Adam is a professional cleaner and has recently joined a new cleaning company that cleans spaces professionally. Soon after he joins, Adam is sent over to a client's house to mop and shine their floors. He is given a bag of the company's self-made supplies and is asked to use only those to do the cleaning. When Adam reaches the client's house, he pays close attention to the floors, and observes that some of them are made with pure hardwood. Adam knows that pure hardwood floors need to be cleaned only with a gentle, chemical-free cleaner. Adam likes his job and wants to make sure to do a good job. He carefully examines his supply bag and picks out the cleaner labelled 'gentle' to do the cleaning.

(Thought: occurrent): When Adam reaches the client's house, he pays close attention to the floors, and observes that some of them are made with pure hardwood. Adam knows that pure hardwood floors need to be cleaned only with a gentle, chemical-free, cleaner. Adam likes his job and wants to make sure to do a good job. He carefully examines his supply bag and picks out the cleaner labelled 'gentle' to do the cleaning. It turns out that the cleaning company's self-made gentle cleaner contains a small but significant percentage of chemical bleach in it a fact not mentioned on the bottle. Adam knows that the gentle cleaner has bleach in it because his boss had told him. However, when he grabs the bottle from his bag, he forgets that the 'gentle' cleaner contains any harsh chemicals, and he decides to use it to clean the floors. However, it turns out that the cleaning company's self-made gentle cleaner contains a small but significant percentage of chemical bleach in it (a fact not mentioned on the bottle). Adam knows that the gentle cleaner has bleach in it because his boss had told him. *(Thought: not occurrent):* However, when Adam grabs the bottle from his bag, he forgets that the 'gentle' cleaner contains some harsh chemicals and decides to use the cleaner on the client's floors. Adam is about to begin cleaning the floors

with the unsafe cleaner when the client suddenly remembers that they have an important event to attend that day and asks Adam to reschedule.

(Outcome): As a result, a different employee returns to the client's house later that week and properly cleans the floors.

B.5.3 Context: Laptop

(*Common background*): Simon is a personal assistant to the CEO of a large multinational corporation. Simon's boss is giving an important presentation later in the day but his charger has stopped working. He asks Simon to find him a laptop charger. Simon admires his boss and likes his work. He cares about the project his boss is working on and wants the presentation to go smoothly. He examines the laptop carefully and selects a spare charger from the office storeroom. He ends up picking out a 100-watt laptop charger. It turns out that Simon's boss's laptop is an old one and can handle only 60 watts of charge at once. Simon knows this because his boss had mentioned it to him once before.

(Thought: occurrent): As Simon grabs the 100-watt charger, he recalls that the laptop's charging capacity. As a result, he puts the 100-watt charger back and select a correct 60-watt charger instead. As Simon opens his boss's laptop to plug in the 60-watt charger, he notices that it is fully charged. His boss ends up having to present sooner than expected. *(Thought: not occurrent):* However, as Simon is selecting a charger, he completely forgets about the laptop's charging capacity. He picks the wrong 100-watt charger to plug into his boss' laptop. As Simon opens his boss's laptop to plug in the 100-watt charger, he notices that it is fully charged.

(Outcome): As a result, the charger never gets plugged into the laptop.

B.5.4 Context: Aquarium

(*Common background*): Mary is a primary schoolteacher, teaching 6-year-olds. Every year, the school takes their 6-year-olds for a field trip to the aquarium. The students have been looking forward to the trip all year. This year, the school's head teacher has tasked Mary with handling the logistics of the trip. The aquarium strictly regulates the number of people inside at any one time. For this reason, tickets are booked in slots and no one is allowed to enter without a ticket. The school's headteacher purchased tickets for all the 6-year-olds months in advance and gave them all to Mary. Mary is looking forward to the trip. On the morning of the field trip, she gets up early and spends time carefully going over the plan for the day and the things she will need.

(Thought: occurrent): As she is leaving her house, she remembers that she needs to pack the tickets for the aquarium. She puts the tickets in her bag and leaves for the aquarium. Mary and the rest of the teachers and students get to the aquarium at their designated time. Mary has the tickets for entry. *(Thought: not occurrent):* However, as she is leaving her house, she fails to realize that she needs to pack the tickets for the aquarium. She leaves for the aquarium without putting the students' tickets in her bag. Mary and the rest of the teachers and students get to the aquarium at their designated time. Mary does not have the tickets for entry.

(Outcome): However, the aquarium had an unexpected incident in the morning and had to shut down temporarily for the day. As a result, the school ends up having to reschedule their visit to the aquarium.

B.5.5 Context: Cake

(*Common background*): Zoey's sister Molly is pregnant and is having a baby shower at her place. The shower is in intimate gathering of Molly's closest friends and family and is a potluck. Molly has asked Zoey to bring the cake. Zoey needs to go to the grocery store to get all the ingredients. Before making that trip, Zoey carefully plans her trip to the store, including making

a detailed list of everything she will need. When she reaches the aisle with the icing, Zoey reaches out for the tube of chocolate icing. It turns out that Molly cannot have chocolate icing. Since the start of the pregnancy, anything chocolate makes Molly severely nauseous. Zoey knows this well because Molly had told her.

(Thought: occurrent): At she is picking the icing, Zoey recalls Molly's intolerance. She puts the tube of chocolate icing back and picks up the tube of vanilla icing instead. She makes the cake with the vanilla icing. On the day of the baby shower, Zoey makes her way to the party with her vanilla iced cake. *(Thought: not occurrent):* However, as she is picking the icing, Zoey forgets about Molly's intolerance and picks up the tube of chocolate icing. She makes the cake with the chocolate icing, not recalling that her sister is intolerant to it. On the day of the baby shower, Zoey makes her way to the party with her chocolate iced cake.

(Outcome): Molly however ends up having to make a last-minute trip to the hospital due to pre-mature labor pains. The shower ends up getting postponed and Zoey goes back home with her vanilla cake.

B.5.6 Context: Gardening

(Common background): Tim is a gardener. He was hired by Anna. Anna has a garden in her backyard that she wants to use to grow organic fruits and vegetables. She has been trying to grow the plants herself but has so far been unsuccessful. She has hired Tim to help her. After speaking to Anna over the phone and accepting the job, Tim goes over to Anna's place. Tim really likes his work and enjoys making clients happy. Tim inspects the garden carefully and notes that the plants will need more fertilizers to grow. He examines the plants and the supplies in his bag to select the best fertilizer for the job. He picks a new fertilizer he has bought recently. The fertilizer is mostly organic but has some strong industrial chemicals. It turns out that Anna wants

Tim to use only organic manures and fertilizers in the garden. Tim knows this because Anna told him this at the time of hiring him.

(Thought: occurrent): After Tim grabs the new fertilizer, he recalls Anna's wish to only use organic fertilizers on her garden. He picks a different fertilizer that is 100% organic instead. He opens the bottle to use the organic fertilizer he selected only to realize that he didn't have the mixing liquid required for it. *(Thought: not occurrent):* However, as Tim is picking a fertilizer, Anna's request completely slips his mind. He picks the new fertilizer that has strong industrial chemicals in it. He opens the bottle to use the fertilizer only to realize that he doesn't have the mixing liquid required for the fertilizer.

(Outcome): He ends up using a different, older fertilizer that is also 100% organic.

B.6 Vignettes used in Experiment 6

B.6.1 Context: Food

(*Common background*): Katherine is an employee at a company. The head of the company asks Katherine to plan an end-of-year dinner to celebrate their accomplishments. To do so, Katherine works with a local restaurant to prepare for this event. The restaurant shows Katherine a menu for her approval. Katherine spends many hours deliberating each dish on the menu. She wants to make sure everyone's dietary restrictions are addressed. She sees that one of the dishes is a pasta dish with some anchovies blended into the tomato sauce for a salty flavor. As it happens, one of the employees, Jan, is quite allergic to anchovies. Katherine knows this, because once Jan had told her. However, as she is deliberating over the food menu, Katherine fails to recall that Jan has an allergy to anchovies. As a result, she approves the pasta dish.

(Exogenous reminder): Later in the day, Katherine and Jan happen to cross each other in the hallway. On seeing her, Jan reminds Katherine of her allergy to anchovies. Immediately after the

encounter, Katherine calls the restaurant and cancels the pasta dish with the anchovy sauce. On the night of the party, the final menu does not include the pasta dish that is made with anchovies to which Jan is allergic. *(No reminder):* Later in the day, the head of the company checks in with Katherine about her progress. After the update, Katherine calls the restaurant to confirm the menu and the catering details. She makes no changes. The final menu includes the pasta dish made with anchovies to which Jan is allergic. However, on the night of the party, the restaurant has to substitute the pasta dish with a salad due to issues with their food stock. As a result, the pasta dish with the anchovies is not part of the dinner.

(Outcome): Jan also has a last-minute change to her plans and she ends up giving the company party a miss.

B.6.2 Context: Cleaning

(*Common background*): Adam is a professional cleaner and has recently joined a new cleaning company that cleans spaces professionally. Soon after he joins, Adam is sent over to a client's house to mop and shine their floors. He is given a bag of the company's self-made supplies and is asked to use only those to do the cleaning. When Adam reaches the client's house, he pays close attention to the floors, and observes that some of them are made with pure hardwood. Adam knows that pure hardwood floors need to be cleaned only with a gentle, chemical-free cleaner. Adam likes his job and wants to make sure to do a good job. He carefully examines his supply bag and picks out the cleaner labelled 'gentle' to do the cleaning. It turns out that the cleaning company's self-made gentle cleaner contains a small but significant percentage of chemical bleach (a fact not mentioned on the bottle). Adam knows that the gentle cleaner has bleach in it because his boss had told him. However, when Adam grabs the bottle

from his bag, he forgets that the ‘gentle’ cleaner contains some harsh chemicals and decides to use the cleaner on the client's floors.

(Exogenous reminder): As Adam grabs the bottle from his bag, he happens to get a phone call from his boss. His boss reminds Adam to not use the ‘gentle’ cleaner on the client’s hardwood floors because it contains harsh chemicals. Adam swaps the bottle out for a safe cleaner instead. Adam is about to begin cleaning the floors with a safe cleaner when the client suddenly remembers that they have an important event to attend that day and asks Adam to reschedule. *(No reminder):* As Adam grabs the bottle from his bag, the client walks into the room to talk to Adam. He double-checks the company’s rates with Adam and Adam confirms that the client’s payment estimation is correct. Adam opens the bottle of the unsafe cleaner to start cleaning. Adam is about to begin cleaning the floors with the unsafe cleaner when the client suddenly remembers that they have an important event to attend that day and asks Adam to reschedule.

(Outcome): As a result, a different employee returns to the client’s house later that week and properly cleans the floors.

B.6.3 Context: Laptop

(*Common background*): Simon is a personal assistant to the CEO of a large multinational corporation. Simon’s boss is giving an important presentation later in the day but his charger has stopped working. He asks Simon to find him a laptop charger. Simon admires his boss and likes his work. He cares about the project his boss is working on and wants the presentation to go smoothly. He examines the laptop carefully and selects a spare charger from the office store room. He ends up picking out a 100-watt laptop charger. It turns out that Simon’s boss’s laptop is an old one and can handle only 60 watts of charge at once. Simon knows this because his boss

had mentioned it to him once before. However, as Simon is selecting a charger, he completely forgets about the laptop's charging capacity. He picks the wrong 100-watt charger to plug into his boss' laptop.

(Exogenous reminder): On his way back to his desk, Simon bumps into one of his coworkers. The coworker sees that Simon is holding a charger and reminds him that their boss' laptop can only handle a 60-watt charger. As a result, Simon goes back and grabs a safe, 60-watt charger instead. As Simon opens his boss's laptop to plug in the 60-watt charger, he notices that it is fully charged. *(No reminder):* Simon bumps into one of his coworkers. The coworker asks Simon whether he would like to get coffee later, but Simon tells him that he'll be busy helping with his boss' presentation. Then Simon continues to his desk with the unsafe, 100-watt charger. As Simon opens his boss's laptop to plug in the 100-watt charger, he notices that it is fully charged.

(Outcome): His boss ends up having to present sooner than expected. As a result, the charger never gets plugged into the laptop.

B.6.4 Context: Aquarium

(*Common background*): Mary is a primary schoolteacher, teaching 6-year-olds. Every year, the school takes their 6-year-olds for a field trip to the aquarium. The students have been looking forward to the trip all year. This year, the school's head teacher has tasked Mary with handling the logistics of the trip. The aquarium strictly regulates the number of people inside at any one time. For this reason, tickets are booked in slots and no one is allowed to enter without a ticket. The school's headteacher purchased tickets for all the 6-year-olds months in advance and gave them all to Mary. Mary is looking forward to the trip. On the morning of the field trip, Mary gets up early and spends time carefully going over the plan for the day and the things she will need.

However, as she is leaving her house, she fails to realize that she needs to pack the tickets for the aquarium. She leaves for the aquarium without putting the students' tickets in her bag.

(Exogenous reminder): As she is pulling her car out of the driveway, Mary receives a text message from the headteacher reminding her not to forget the tickets. Mary rushes back inside and grabs the tickets, placing them in her bag. Mary and the rest of the teachers and students get to the aquarium at their designated time. Mary has the tickets for entry. *(No reminder):* As she is pulling her car out of the driveway, Mary receives a text message from her brother asking her what she is up to during the day. Mary sends him a quick summary of her plans for the day before continuing to the aquarium. She drives away from her house without the tickets. Mary and the rest of the teachers and students get to the aquarium at their designated time. Mary does not have the tickets for entry.

(Outcome): However, the aquarium had an unexpected incident in the morning and had to shut down temporarily for the day. As a result, the school ends up having to reschedule their visit to the aquarium.

B.6.5 Context: Cake

(*Common background*): Zoey's sister Molly is pregnant and is having a baby shower at her place. The shower is in intimate gathering of Molly's closest friends and family and is a potluck. Molly has asked Zoey to bring the cake. Zoey needs to go to the grocery store to get all the ingredients. Before making that trip, Zoey carefully plans her trip to the store, including making a detailed list of everything she will need. When she reaches the aisle with the icing, Zoey reaches out for the tube of chocolate icing. It turns out that Molly cannot have chocolate icing. Since the start of the pregnancy, anything chocolate makes Molly severely nauseous. Zoey knows this well because Molly had told her. However, as she is picking the icing, Zoey forgets

about Molly's intolerance and picks up the tube of chocolate icing. She makes the cake with the chocolate icing, not recalling that her sister is intolerant to it.

(Exogenous reminder): The evening before the baby shower, Zoey receives a phone call from her mother. Her mother reminds Zoey that Molly is intolerant to chocolate. As a result, Zoey quickly makes another cake with vanilla icing she had lying around the house. On the day of the baby shower, Zoey makes her way to the party with her vanilla iced cake. *(No reminder):* The evening before the baby shower, Zoey receives a phone call from her mother confirming the details of the party the next day. Zoey confirms that she will be attending and bringing a cake. After the call, Zoey adds finishing touches to the chocolate cake she plans to take to the shower. On the day of the baby shower, Zoey makes her way to the party with her chocolate iced cake.

(Outcome): Molly however ends up having to make a last-minute trip to the hospital due to pre-mature labor pains. The shower ends up getting postponed and Zoey goes back home with her vanilla cake.

B.6.6 Context: Gardening

(*Common background*): Tim is a gardener. He was hired by Anna. Anna has a garden in her backyard that she wants to use to grow organic fruits and vegetables. She has been trying to grow the plants herself but has so far been unsuccessful. She has hired Tim to help her. After speaking to Anna over the phone and accepting the job, Tim goes over to Anna's place. Tim really likes his work and enjoys making clients happy. Tim inspects the garden carefully and notes that the plants will need more fertilizers to grow. He examines the plants and the supplies in his bag to select the best fertilizer for the job. He picks a new fertilizer he has bought recently. The fertilizer is mostly organic but has some strong chemicals. It turns out that Anna wants Tim to use only organic manures and fertilizers in the garden. Tim knows this because Anna told him

this at the time of hiring him. However, as Tim is picking a fertilizer, Anna's request completely slips his mind. He picks the new fertilizer that has strong chemicals in it.

(Exogenous reminder): After selecting the fertilizer, Anna talks to Tim and reminds him that she only wants him to use organic fertilizers on her garden. As a result, Tim switches the fertilizer with strong chemicals out for an old fertilizer he had purchased a while back. This other fertilizer is completely organic. He opens the bottle to use the organic fertilizer he selected only to realize that he didn't have the mixing liquid required for it. *(No reminder):* After selecting the fertilizer, Anna talks to Tim to confirm the details of his services. Tim confirms the times and rates they had previously agreed on. He picks up the fertilizer with the strong chemicals to start his work. He opens the bottle to use the fertilizer only to realize that he doesn't have the mixing liquid required for the fertilizer.

(Outcome): He ends up using a different, older fertilizer that is also 100% organic.

References

Afifi, T. O., Ford, D., Gershoff, E. T., Merrick, M., Grogan-Kaylor, A., Ports, K. A., MacMillan, H. L., Holden, G. W., Taylor, C. A., Lee, S. J., & Peters Bennett, R. (2017). Spanking and adult mental health impairment: The case for the designation of spanking as an adverse childhood experience. *Child Abuse & Neglect*, *71*, 24–31. https://doi.org/10.1016/j.chiabu.2017.01.014

Alicke, M. (2000). Culpable Control and the Psychology of Blame. *Psychological Bulletin*, *126*, 556–574. https://doi.org/10.1037//0033-2909.126.4.556

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310. https://doi.org/10.1073/pnas.1912341117

Amaya, S. (2022). Negligence: Its Moral Significance. In M. Vargas & J. M. Doris (Eds.), *The Oxford Handbook of Moral Psychology* (pp. 660–682). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198871712.013.33

Amaya, S., & Doris, J. M. (2015). No Excuses: Performance Mistakes in Morality. In J. Clausen & N. Levy (Eds.), *Handbook of Neuroethics* (pp. 253–272). Springer.

Arpaly, N., & Schroeder, T. (1999). Praise, Blame and the Whole Self. *Philosophical Studies*, *93*, 161–188.

Ashley, J., & Tomasello, M. (1998). Cooperative Problem-Solving and Teaching in Preschoolers. *Social Development*, *7*(2), 143–163. https://doi.org/10.1111/1467-9507.00059

Azrin, N. H., & Holz, W. C. (1966). Punishment. In *Operant behavior* (W. K. Honig (Ed.)). Appleton-Century-Crofts.

REFERENCES

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), Article 4. https://doi.org/10.1038/s41562-017-0064

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(3).

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters*, *122*(2), 308–310. https://doi.org/10.1016/j.econlet.2013.11.028

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111*(45), 15924–15927. https://doi.org/10.1073/pnas.1413170111

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms13327

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences*, *108*(27), 11023–11027. https://doi.org/10.1073/pnas.1105456108

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference*. *Journal of Child Language*, *20*(2), 395–418. https://doi.org/10.1017/S0305000900008345

REFERENCES

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Bandura, A. (1962). Social learning through imitation. In *Nebraska Symposium on Motivation, 1962* (pp. 211–274). Univer. Nebraska Press.

Bandura, A. (1977). *Social learning theory* (pp. viii, 247). Prentice-Hall.

Barker, G. P., & Graham, S. (1987). Developmental study of praise and blame as attributional cues. *Journal of Educational Psychology*, *79*, 62–66. https://doi.org/10.1037/0022-0663.79.1.62

Baron, J., & Hershey, J. C. (1988). Outcome Bias in Decision Evaluation. *Journal of Personality and Social Psychology*, *54*, 11.

Baron-Cohen, S. (2012). Evolution of a theory of mind? *The Descent of Mind: Psychological Perspectives on Hominid Evolution*. https://doi.org/10.1093/acprof:oso/9780192632593.003.0013

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barrett, A., & Harrison, C. (2005). *Crime and Punishment in England: A Sourcebook*. Routledge.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspectives*

on Psychological Science, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, *194*, 104057. https://doi.org/10.1016/j.cognition.2019.104057

Bear, A., & Knobe, J. (2017). *Normality: Part descriptive, part prescriptive*. 13.

Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, *60*(6), 503–515. https://doi.org/10.1090/S0002-9904-1954-09848-8

Berg-Cross, L. G. (1975). Intentionality, Degree of Damage, and Moral Judgments. *Child Development*, *46*(4), 970–974. https://doi.org/10.2307/1128406

Bernhard, R. M., Martin, J. W., & Warneken, F. (2020). Why do children punish? Fair outcomes matter more than intent in children's second- and third-party punishment. *Journal of Experimental Child Psychology*, *200*, 104909. https://doi.org/10.1016/j.jecp.2020.104909

Berns, G. S., Laibson, D., & Loewenstein, G. (2007). Intertemporal choice – toward an integrative framework. *Trends in Cognitive Sciences*, *11*(11), 482–488. https://doi.org/10.1016/j.tics.2007.08.011

Bloom, P. (2002). *How Children Learn the Meanings of Words*. MIT Press.

Boehm, C. (2001). Hierarchy in the Forest: The Evolution of Egalitarian Behavior. In *Hierarchy in the Forest*. Harvard University Press. https://doi.org/10.4159/9780674028449

Boehm, C. (2011). Retaliatory Violence in Human Prehistory. *British Journal of Criminology*, *51*, 518.

REFERENCES

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330. https://doi.org/10.1016/j.cognition.2010.10.001

Boon, S. D., Deveau, V. L., & Alibhai, A. M. (2009). Payback: The parameters of revenge in romantic relationships. *Journal of Social and Personal Relationships*, *26*(6–7), 747–768. https://doi.org/10.1177/0265407509347926

Bornstein, A. M., & Daw, N. D. (2011). Multiplicity of control in the basal ganglia: Computational roles of striatal subregions. *Current Opinion in Neurobiology*, *21*(3), 374–380. https://doi.org/10.1016/j.conb.2011.02.009

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130480. https://doi.org/10.1098/rstb.2013.0480

Bourgeois-Gironde, S. (2010). Regret and the rationality of choices. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1538), 249–257. https://doi.org/10.1098/rstb.2009.0163

Bowles, S., & Gintis, H. (2013). A Cooperative Species. *Introductory Chapters*. https://ideas.repec.org//h/pup/chapts/9474-1.html

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, *100*(6), 3531–3535. https://doi.org/10.1073/pnas.0630443100

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*(3), 171–195. https://doi.org/10.1016/0162-3095(92)90032-Y

REFERENCES

Bradac, J. J. (2001). Theory Comparison: Uncertainty Reduction, Problematic Integration, Uncertainty Management, and Other Curious Constructs. *Journal of Communication*, *51*(3), 456–476. https://doi.org/10.1111/j.1460-2466.2001.tb02891.x

Bradford, W. D., Dolan, P., & Galizzi, M. M. (2019). Looking ahead: Subjective time perception and individual discounting. *Journal of Risk and Uncertainty*, *58*(1), 43–69. https://doi.org/10.1007/s11166-019-09298-1

Brandt, H., Hauert, C., & Sigmund, K. (2003). Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1519), 1099–1104. https://doi.org/10.1098/rspb.2003.2336

Brehm, J. W. (1966). *A theory of psychological reactance* (pp. x, 135). Academic Press.

Brown, P., Levinson, S. C., & Gumperz, J. J. (1987, February 26). *Politeness: Some Universals in Language Usage*. Higher Education from Cambridge University Press; Cambridge University Press. https://doi.org/10.1017/CBO9780511813085

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Burnham, M., Le, Y., & Piedmont, R. (2018). Who is Mturk? Personal characteristics and sample consistency of these online workers*. In *Mental Health, Religion & Culture* (Vol. 21). https://doi.org/10.1080/13674676.2018.1486394

Carey, A. A., Tanaka, Y., & van Der Meer, M., A. (2019). Reward revaluation biases hippocampal replay content away from the preferred outcome. *Nature Neuroscience*, *22*(9), 1450–1459.

REFERENCES

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437–451. https://doi.org/10.1016/j.jesp.2005.06.007

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*, 284–299. https://doi.org/10.1037/0022-3514.83.2.284

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, *95*(6), 1316–1324. https://doi.org/10.1037/a0012165

Cesana-Arlotti, N., Téglás, E., & Bonatti, L. (2012). The Probable and the Possible at 12 Months: Intuitive Reasoning about the Uncertain Future. *Advances in Child Development and Behavior*, *43*, 1–25.

Chambliss, W. J. (1966). The Deterrent Influence of Punishment. *Crime and Delinquency*, *12*, 70.

Chandrasekhar, P. V. S., Capra, C. M., Moore, S., Noussair, C., & Berns, G. S. (2008). Neurobiological regret and rejoice functions for aversive outcomes. *NeuroImage*, *39*(3), 1472–1484. https://doi.org/10.1016/j.neuroimage.2007.10.027

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*(1), 1–15. https://doi.org/10.1017/S0140525X00001515

Clarke, R. (2014a). Negligent action and unwitting omission. In *Surrounding Free Will: Philosophy, Psychology, Neuroscience* (pp. 298–317).

Clarke, R. (2014b). Omissions: Agency, metaphysics, and responsibility. In *Omissions: Agency, Metaphysics, and Responsibility*. Oxford University Press.

REFERENCES

Clarke, R. (2017). Blameworthiness and unwitting omissions. In *The ethics and law of omissions* (pp. 63–83).

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*(6511), 209–216. https://doi.org/10.1038/373209a0

Collins, J. (2006). Crime and parenthood: The uneasy case for prosecuting negligent parents. *Nw. UL REv.*, *100*, 807.

Colston, H. L., & O'Brien, J. (2000). Contrast of Kind Versus Contrast of Magnitude: The Pragmatic Accomplishments of Irony and Hyperbole. *Discourse Processes*, *30*(2), 179–199. https://doi.org/10.1207/S15326950DP3002_05

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: A neuroimaging study of choice behavior. *Nature Neuroscience*, *8*(9), Article 9. https://doi.org/10.1038/nn1514

Critcher, C. R., Helzer, E. G., Tannenbaum, D., & Pizarro, D. A. (2017). *Moral evaluations depend upon mindreading moral occurrent beliefs*.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How Quick Decisions Illuminate Moral Character. *Social Psychological and Personality Science*, *4*(3), 308–315. https://doi.org/10.1177/1948550612457688

Cruciani, F., Berardi, A., Cabib, S., & Conversi, D. (2011). Positive and negative emotional arousal increases duration of memory traces: Common and independent mechanisms. *Frontiers in Behavioral Neuroscience*, *5*. https://www.frontiersin.org/articles/10.3389/fnbeh.2011.00086

Csibra, G. (2007). Teachers in the wild. *Trends in Cognitive Sciences*, *11*(3), 95–96. https://doi.org/10.1016/j.tics.2006.12.001

REFERENCES

Csibra, G., & Gergely, G. (2006). *Social learning and social cognition: The case for pedagogy*.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F. (2013). Action, Outcome, and Value: A Dual-System Framework for Morality. *Personality and Social Psychology Review*, *17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Cushman, F. (2015). Punishment in Humans: From Intuitions to Institutions: Punishment: Intuitions and Institutions. *Philosophy Compass*, *10*(2), 117–133. https://doi.org/10.1111/phc3.12192

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental Outcomes Guide Punishment in a "Trembling Hand" Game. *PLOS ONE*, *4*(8), e6699. https://doi.org/10.1371/journal.pone.0006699

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, *112*(45), 13817–13822. https://doi.org/10.1073/pnas.1506367112

Cushman, F., Sarin, A., & Ho, M. K. (2019). *Punishment as communication*. PsyArXiv. https://doi.org/10.31234/osf.io/wf3tz

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. https://doi.org/10.1016/j.cognition.2012.11.008

REFERENCES

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204. https://doi.org/10.1016/j.conb.2006.03.006

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. https://doi.org/10.1016/j.neuron.2011.02.027

Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*, *22*(6), 1068–1074. https://doi.org/10.1016/j.conb.2012.05.011

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268.

Delgado, M., Jou, R., LeDoux, J., & Phelps, L. (2009). Avoiding negative outcomes: Tracking the mechanisms of avoidance learning in humans during fear conditioning. *Frontiers in Behavioral Neuroscience*, *3*. https://www.frontiersin.org/articles/10.3389/neuro.08.033.2009

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, *108*(32), 13335–13340. https://doi.org/10.1073/pnas.1102131108

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

REFERENCES

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated Moral Reasoning. In *Psychology of Learning and Motivation* (Vol. 50, pp. 307–338). Elsevier. https://doi.org/10.1016/S0079-7421(08)00410-6

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410–416. https://doi.org/10.1038/nn2077

Dugar, S. (2010). Nonmonetary sanctions and rewards in an experimental coordination game. *Journal of Economic Behavior & Organization*, *73*(3), 377–386. https://doi.org/10.1016/j.jebo.2009.11.003

Durkheim, E. (2014). The Division of Labor in Society. In *Social Stratification* (4th ed.). Routledge.

Epstein, L. G. (1999). A Definition of Uncertainty aversion. *The Review of Economic Studies*, *66*(3), 579–608.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, *53*(5), 495–511. https://doi.org/10.1016/j.euroecorev.2008.09.007

Ewing, A. C. (1943). Punishment as Viewed by the Philosopher. *Canadian Bar Review*, *21*, 102.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 4.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*(6928), Article 6928. https://doi.org/10.1038/nature01474

Feinberg, J. (1965). The Expressive Function of Punishment. *The Monist*, *49*(3), 397–423.

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences*, *25*(12), 1045–1057. https://doi.org/10.1016/j.tics.2021.09.002

Fincham, F. D. (1982). Responsibility Attribution in the Culturally Deprived. *The Journal of Genetic Psychology*, *140*(2), 229–235. https://doi.org/10.1080/00221325.1982.10534195

Fincham, F. D., & Jaspars, J. M. (1980). Attribution of Responsibility: From Man the Scientist to Man As Lawyer. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 13, pp. 81–138). Academic Press. https://doi.org/10.1016/S0065-2601(08)60131-8

Fischhoff, B. (2003). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Quality & Safety in Health Care*, *12*, 304–311; discussion 311. https://doi.org/10.1037/0096-1523.1.3.288

Fox, K. R., Toole, K. E., Franklin, J. C., & Hooley, J. M. (2017). Why Does Nonsuicidal Self-Injury Improve Mood? A Preliminary Test of Three Hypotheses. *Clinical Psychological Science*, *5*(1), 111–121. https://doi.org/10.1177/2167702616662270

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160. https://doi.org/10.3758/CABN.1.2.137

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *Vol. 68*(No. 1), 5–20.

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology: CB*, *15*(17), R644-646. https://doi.org/10.1016/j.cub.2005.08.041

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the Message: Punishment Is Satisfying If the Transgressor Responds to Its Communicative Intent. *Personality and Social Psychology Bulletin*, *40*(8), 986–997. https://doi.org/10.1177/0146167214533130

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gergely, G., & Csibra, G. (2005). The social construction of the cultural mind: Imitative learning as a mechanism of human pedagogy. *Interaction Studies*, *6*(3), 463–481. https://doi.org/10.1075/is.6.3.10ger

Gershman, S. J., Zhou, J., & Kommers, C. (2017). Imaginative Reinforcement Learning: Computational Principles and Neural Mechanisms. *Journal of Cognitive Neuroscience*, *29*(12), 2103–2113. https://doi.org/10.1162/jocn_a_01170

Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin*, *128*(4), 539–579. https://doi.org/10.1037/0033-2909.128.4.539

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Glassman, L. H., Weierich, M. R., Hooley, J. M., Deliberto, T. L., & Nock, M. K. (2007). Child maltreatment, non-suicidal self-injury, and the mediating role of self-criticism. *Behaviour Research and Therapy*, *45*(10), 2483–2490. https://doi.org/10.1016/j.brat.2007.04.002

Glimcher, P., & Fehr, E. (2014). Neuroeconomics: Decision making and the brain. In *Neuroeconomics: Decision making and the brain. Edited by: Glimcher, Paul W; Fehr, Ernst (2014). London: Academic Press Inc.* (pp. 171–186). Academic Press Inc. https://www.zora.uzh.ch/id/eprint/84485/

Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology*, *45*(4), 840–844. https://doi.org/10.1016/j.jesp.2009.03.001

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, *41*(3), 364–374. https://doi.org/10.1002/ejsp.782

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184. https://doi.org/10.1111/tops.12007

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. MIT Press.

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Graybiel, A. M. (2008). Habits, Rituals, and the Evaluative Brain. *Annual Review of Neuroscience*, *31*(1), 359–387. https://doi.org/10.1146/annurev.neuro.29.051605.112851

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388. https://doi.org/10.2307/2182440

Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the Leviathan – Voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep20767

REFERENCES

Gruber, A. J., Dayan, P., Gutkin, B. S., & Solla, S. A. (2006). Dopamine modulation in the basal ganglia locks the gate to working memory. *Journal of Computational Neuroscience*, *20*(2), 153–166. https://doi.org/10.1007/s10827-005-5705-x

Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage*, *62*(1), 154–166. https://doi.org/10.1016/j.neuroimage.2012.04.024

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The Competitive Advantage of Sanctioning Institutions. *Science (New York, N.Y.)*, *312*, 108–111. https://doi.org/10.1126/science.1123633

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910. https://doi.org/10.1016/j.tics.2021.07.008

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Hampton, J. (1984). The Moral Education Theory of Punishment. *Philosophy & Public Affairs*, *13*(3), 208–238.

Hart, H. L. (1968). Negligence, mens rea, and criminal responsibility. In *Punishment and Responsibility: Essays in the Philosophy of Law* (Second Edition, pp. 136–157). Oxford University Press.

Hartman, R. J. (2019). Moral luck and the unfairness of morality. *Philosophical Studies*, *176*(12), 3179–3197. https://doi.org/10.1007/s11098-018-1169-5

REFERENCES

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research*, *67*(8), 1688–1699. https://doi.org/10.1016/j.jbusres.2014.02.015

Hauser, J. R., & Wernerfelt, B. (1990). An Evaluation Cost Model of Consideration Sets. *Journal of Consumer Research*, *16*(4), 393. https://doi.org/10.1086/209225

Heider, F. (1958). *The psychology of interpersonal relations* (pp. ix, 326). John Wiley & Sons Inc. https://doi.org/10.1037/10628-000

Henrich, J. (2006). Cooperation, Punishment, and the Evolution of Human Institutions. *Science*, *312*(5770), 60–61. https://doi.org/10.1126/science.1126398

Henrich, J. (2015). The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter. In *The Secret of Our Success*. Princeton University Press. https://doi.org/10.1515/9781400873296

Henrich, J., & Boyd, R. (2001). Why People Punish Defectors. *Journal of Theoretical Biology*, *208*(1), 79–89. https://doi.org/10.1006/jtbi.2000.2202

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly Punishment Across Human Societies. *Science*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis. *Science*, *317*(5843), 1360–1366. https://doi.org/10.1126/science.1146282

Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *The Journal of Philosophy*, 587–612.

REFERENCES

Ho, M. K., Cushman, F., Littman, M., & Austerweil, J. L. (2018). *People Teach with Rewards and Punishments as Communication not Reinforcements*. PsyArXiv. https://doi.org/10.31234/osf.io/3cd7r

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology. General*, *150*(11), 2246–2272. https://doi.org/10.1037/xge0001035

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/b5488aeff42889188d03c9895255cecc-Abstract.html

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91–106. https://doi.org/10.1016/j.cognition.2017.03.006

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with Theory of Mind. *Trends in Cognitive Sciences*, *26*(11), 959–971. https://doi.org/10.1016/j.tics.2022.08.003

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral Punishment in Everyday Life. *Personality and Social Psychology Bulletin*, *44*(12), 1697–1711. https://doi.org/10.1177/0146167218775075

Hooley, J. M., & St. Germain, S. A. (2014). Nonsuicidal Self-Injury, Pain, and Self-Criticism: Does Changing Self-Worth Change Pain Endurance in People Who Engage in Self-Injury? *Clinical Psychological Science*, *2*(3), 297–305. https://doi.org/10.1177/2167702613509372

REFERENCES

Houser, D., Xiao, E., McCabe, K., & Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior*, *62*(2), 509–532. https://doi.org/10.1016/j.geb.2007.05.001

Howard, J. A., & Sheth, J. N. (1969). *The theory of buyer behavior* (pp. xv, 458). John Wiley & Sons.

Isbell, C., & Shelton, C. (2001). Cobot: A Social Reinforcement Learning Agent. *Advances in Neural Information Processing Systems*, *14*. https://proceedings.neurips.cc/paper/2001/hash/92bbd31f8e0e43a7da8a6295b251725f-Abstract.html

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110. https://doi.org/10.1016/j.cobeha.2019.04.010

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334. https://doi.org/10.1016/j.cogpsych.2020.101334

Johnson, J. G., & Raab, M. (2003). Take The First: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes*, *91*(2), 215–229. https://doi.org/10.1016/S0749-5978(03)00027-X

REFERENCES

Jong, N. K., & Stone, P. (2008). *Hierarchical Model-Based Reinforcement Learning: R-MAX + MAXQ*. 8.

Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*(1), 57–88. https://doi.org/10.1037/pspi0000186

Joyner, B., & Beaver, K. M. (2022). Unpacking the Association Between Corporal Punishment and Criminal Involvement. *Criminal Justice and Behavior*, *49*(12), 1845–1863. https://doi.org/10.1177/00938548221107874

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633. https://doi.org/10.1038/nn2007

Kaiser, S., Simon, J. J., Kalis, A., Schweizer, S., Tobler, P. N., & Mojzisch, A. (2013). The cognitive and neural basis of option generation and subsequent choice. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 814–829. https://doi.org/10.3758/s13415-013-0175-5

Kalis, A., Kaiser, S., & Mojzisch, A. (2013). Why we should talk about option generation in decision-making research. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00555

Kao, J. T., & Goodman, N. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society*.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007. https://doi.org/10.1073/pnas.1407479111

Keller, L. R., & Ho, J. L. (1988). Decision problem structuring: Generating options. *IEEE Transactions on Systems, Man, and Cybernetics*, *18*(5), 715–728. https://doi.org/10.1109/21.21599

Khoury, A. C. (2018). The objects of moral responsibility. *Philosophical Studies*, *175*(6), 1357–1381. https://doi.org/10.1007/s11098-017-0914-5

King, M. (2009). The problem with negligence. *Social Theory and Practice, 35*(4), 577–595.

Klein, G. A. (Ed.). (1993). *Decision making in action: Models and methods*. Ablex Pub.

Klein, G., Wolf, S., Militello, L., & Zsambok, C. (1995). Characteristics of Skilled Option Generation in Chess. *Organizational Behavior and Human Decision Processes*, *62*(1), 63–69.

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003

Knudsen, B., & Liszkowski, U. (2013). One-Year-Olds Warn Others About Negative Action Outcomes. *Journal of Cognition and Development*, *14*(3), 424–436. https://doi.org/10.1080/15248372.2012.689387

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed Neural Representation of Expected Value. *Journal of Neuroscience*, *25*(19), 4806–4812. https://doi.org/10.1523/JNEUROSCI.0642-05.2005

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking Under the Hood of Third-Party Punishment Reveals Design for Personal Benefit. *Psychological Science*, *27*(3), 405–418. https://doi.org/10.1177/0956797615624469

Kubanek, J., Snyder, L. H., & Abrams, R. A. (2015). Reward and punishment act as distinct factors in guiding behavior. *Cognition*, *139*, 154–167. https://doi.org/10.1016/j.cognition.2015.03.005

Kumle, L., Vo, M. L.-H., & Draschkow, D. (2023). *Mixedpower: A library for estimating simulation-based power for mixed models in R* (R package version 0.1.0).

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770. https://doi.org/10.1016/j.cognition.2008.06.009

Laurent, S. M., Clark, B. A. M., Walker, S., & Wiseman, K. D. (2014). Punishing hypocrisy: The roles of hypocrisy and moral emotions in deciding culpability and punishment of criminal and civil moral transgressors. *Cognition and Emotion*, *28*(1), 59–83. https://doi.org/10.1080/02699931.2013.801339

Laurent, S. M., Nuñez, N. L., & Schweitzer, K. A. (2016). Unintended, but still blameworthy: The roles of awareness, desire, and anger in negligence, restitution, and punishment. *Cognition and Emotion*, *30*(7), 1271–1288. https://doi.org/10.1080/02699931.2015.1058242

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807. https://doi.org/10.1037/a0019688

Lee, R. B. (1979). *The! Kung San: Men, women and work in a foraging society*. Cambridge University Press.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). *emmeans: Estimated marginal means, aka least-squares means* (1.4).

Lerner, M. J., & Clayton, S. (2011). *Justice and Self-Interest: Two Fundamental Motives*. Cambridge University Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged Frequency of Lethal Events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551.

Lieder, F., Griffiths, T., & Hsu, M. (2015). *Utility-weighted sampling in decisions from experience*. https://doi.org/10.13140/RG.2.2.19690.24005

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Over-representation of extreme events in decision-making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1–32. https://doi.org/10.1037/rev0000074

Lin, A., Adolphs, R., & Rangel, A. (2012). Social and monetary reward learning engage overlapping neural substrates. *Social Cognitive and Affective Neuroscience*, *7*(3), 274–281. https://doi.org/10.1093/scan/nsr006

Loewenstein, G., & Thaler, R. H. (1989). Anomalies: Intertemporal Choice. *Journal of Economic Perspectives*, *3*(4), 181–193. https://doi.org/10.1257/jep.3.4.181

Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences*, *104*(22), 9493–9498. https://doi.org/10.1073/pnas.0608842104

Lombrozo, T. (2020). "Learning by Thinking" in Science and in Everyday Life. In T. Lombrozo, *The Scientific Imagination* (pp. 230–249). Oxford University Press. https://doi.org/10.1093/oso/9780190212308.003.0010

Lopez-Brau, M., & Jara-Ettinger, J. (2020). *Physical pragmatics: Inferring the social meaning of object*. https://doi.org/10.31234/osf.io/mnf4y

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215. https://doi.org/10.3758/s13423-015-0808-5

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme Outcomes Sway Risky Decisions from Experience: Risky Decisions and Extreme Outcomes. *Journal of Behavioral Decision Making*, *27*(2), 146–156. https://doi.org/10.1002/bdm.1792

Lutz, K., & Widmer, M. (2014). What can the monetary incentive delay task tell us about the neural processing of reward and punishment? *Neuroscience and Neuroeconomics*, 33. https://doi.org/10.2147/NAN.S38864

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*(3), 629–636. https://doi.org/10.3758/s13423-013-0542-9

Maguire-Jack, K., Gromoske, A. N., & Berger, L. M. (2012). Spanking and Child Development During the First 5 Years of Life. *Child Development*, *83*(6), 1960–1977. https://doi.org/10.1111/j.1467-8624.2012.01820.x

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* MIT Press.

Malle, B. F. (2011). Attribution Theories: How People Make Sense of Behavior. In *Theories in social psychology* (pp. 72–95). Wiley-Blackwell Malden.

Margoni, F., & Surian, L. (2021). Judging accidental harm: Due care and foreseeability of side effects. *Current Psychology*. https://doi.org/10.1007/s12144-020-01334-7

Martin, J. W., & Cushman, F. (2015). To Punish or to Leave: Distinct Cognitive Processes Underlie Partner Control and Partner Choice Behaviors. *PLOS ONE*, *10*(4), e0125193. https://doi.org/10.1371/journal.pone.0125193

Martin, J. W., & Cushman, F. (2016). The Adaptive Logic of Moral Luck. In J. Sytsma & W. Buckwalter (Eds.), *A Companion to Experimental Philosophy* (pp. 190–202). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118661666.ch12

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), Article 11. https://doi.org/10.1038/s41593-018-0232-z

Mazzocco, P., Alicke, M., & Davis, T. (2004). On the Robustness of Outcome Bias: No Constraint by Prior Culpability. *Basic and Applied Social Psychology - BASIC APPL SOC PSYCHOL*, *26*, 131–146. https://doi.org/10.1207/s15324834basp2602&3_3

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160

Meyer, W.-U. (1982). Indirect communications about perceived ability estimates. *Journal of Educational Psychology*, *74*, 888–897. https://doi.org/10.1037/0022-0663.74.6.888

Meyer, W.-U. (1992). Paradoxical Effects of Praise and Criticism on Perceived Ability. *European Review of Social Psychology*, *3*(1), 259–283. https://doi.org/10.1080/14792779243000087

Miller, A., & Perry, R. (2012). The Reasonable Person. *NYUL Rev*, *87*, 323.

Miller, A. T., & Hom, H. L. Jr. (1997). Conceptions of ability and the interpretation of praise, blame and material rewards. *Journal of Experimental Education*, *65*, 163–177. https://doi.org/10.1080/00220973.1997.9943790

Miller, D. T. (2001). Disrespect and the Experience of Injustice. *Annual Review of Psychology*, *52*(1), 527–553. https://doi.org/10.1146/annurev.psych.52.1.527

Mitchell, T. R., & Kalb, L. S. (1981). Effects of outcome knowledge and outcome valence on supervisors' evaluations. *Journal of Applied Psychology*, *66*, 604–612. https://doi.org/10.1037/0021-9010.66.5.604

Molnar, A., Chaudhry, S., & Loewenstein, G. (2022). "It's Not About the Money. It's About Sending a Message!": Avengers Want Offenders to Understand the Reason for Revenge. In *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3524910

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *ELife*, *7*, e32548. https://doi.org/10.7554/eLife.32548

Moore, C., & Pierce, L. (2016). Reactance to Transgressors: Why Authorities Deliver Harsher Penalties When the Social Context Elicits Expectations of Leniency. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00550

Morris, A., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Evolution of flexibility and rigidity in retaliatory punishment. *Proceedings of the National Academy of Sciences*, *114*(39), 10396–10401. https://doi.org/10.1073/pnas.1704032114

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2019). Habits of thought generate candidate actions for choice. *PsyArXiv*.

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating Options and Choosing Between Them Depend on Distinct Forms of Value Representation. *Psychological Science*, *32*(11), 1731–1746. https://doi.org/10.1177/09567976211005702

Morris, H. (1981). A Paternalistic Theory of Punishment. *American Philosophical Quarterly*, *18*(4), 263–271.

Murray, S., Krasich, K., Irving, Z., Nadelhoffer, T., & De Brigard, F. (2022). Mental control and attributions of blame for negligent wrongdoing. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001262

Murray, S., Murray, E. D., Stewart, G., Sinnott-Armstrong, W., & De Brigard, F. (2019). Responsibility for forgetting. *Philosophical Studies*, *176*(5), 1177–1201. https://doi.org/10.1007/s11098-018-1053-3

Murray, S., & Vargas, M. (2020). Vigilance and control. *Philosophical Studies*, *177*(3), 825–843. https://doi.org/10.1007/s11098-018-1208-2

Musculus, L., Ruggeri, A., Raab, M., & Lobinger, B. (2019). A developmental perspective on option generation and selection. *Developmental Psychology*, *55*(4), 745–753. https://doi.org/10.1037/dev0000665

Nadelhoffer, T., Heshmati, S., Kaplan, D., & Nichols, S. (2013). Folk Retributivism and the Communication Confound. *Economics and Philosophy*, *29*, 235–261. https://doi.org/10.1017/S0266267113000217

Nagel, T. (1979). Moral luck. In *Moral Questions* (pp. 31–32). [New York: Cambridge University Press].

Nelkin, D. K. (2004). *Moral Luck*. http://seop.illc.uva.nl/entries/moral-luck/

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151–166. https://doi.org/10.1037/h0048495

Nobes, G., & Martin, J. W. (2022). They should have known better: The roles of negligence and outcome in moral judgements of accidental actions. *British Journal of Psychology*, *113*(2), 370–395. https://doi.org/10.1111/bjop.12536

Nobes, G., Panagiotaki, G., & Pawson, C. (2009). The influence of negligence, intention, and outcome on children's moral judgments. *Journal of Experimental Child Psychology*, *104*, 382–397. https://doi.org/10.1016/j.jecp.2009.08.001

Nock, M. K. (2008). Actions speak louder than words: An elaborated theoretical model of the social functions of self-injury and other harmful behaviors. *Applied and Preventive Psychology*, *12*(4), 159–168. https://doi.org/10.1016/j.appsy.2008.05.002

Nock, M. K., & Prinstein, M. J. (2004). A Functional Approach to the Assessment of Self-Mutilative Behavior. *Journal of Consulting and Clinical Psychology*, *72*(5), 885–890. https://doi.org/10.1037/0022-006X.72.5.885

Norman, D. A., & Shallice, T. (1986). Attention to Action. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and Self-Regulation: Advances in Research and Theory Volume 4* (pp. 1–18). Springer US. https://doi.org/10.1007/978-1-4757-0629-1_1

Nuñez, N., Laurent, S., & Gray, J. M. (2014). Is Negligence a First Cousin to Intentionality? Lay Conceptions of Negligence and Its Relationship to Intentionality. *Applied Cognitive Psychology*, *28*(1), 55–65. https://doi.org/10.1002/acp.2957

Oaksford, M., & Chater, N. (2009). Précis of *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. *Behavioral and Brain Sciences*, *32*(1), 69–84. https://doi.org/10.1017/S0140525X09000284

Olsson, A., & Phelps, E. A. (2004). Learned Fear of "Unseen" Faces after Pavlovian, Observational, and Instructed Fear. *Psychological Science*, *15*(12), 822–828. https://doi.org/10.1111/j.0956-7976.2004.00762.x

O'Reilly, R., & Frank, M. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, *18*, 283–328. https://doi.org/10.1162/089976606775093909

Owen, D. J., Slep, A. M. S., & Heyman, R. E. (2012). The Effect of Praise, Positive Nonverbal Response, Reprimand, and Negative Nonverbal Response on Child Compliance: A Systematic Review. *Clinical Child and Family Psychology Review*, *15*(4), 364–385. https://doi.org/10.1007/s10567-012-0120-0

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Patil, I. (2021). Visualizations with statistical details: The "ggstatsplot" approach. *Journal of Open Source Software*, *6*(61), 3167. https://doi.org/10.21105/joss.03167

Perner, J. (1991). *Understanding the representational mind* (pp. xiv, 348). The MIT Press.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical Active Inference: A Theory of Motivated Control. *Trends in Cognitive Sciences*, *22*(4), 294–306. https://doi.org/10.1016/j.tics.2018.01.009

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*, 201619717. https://doi.org/10.1073/pnas.1619717114

Phillips, J., Morris, A., & Cushman, F. (2019). How We Know What Not To Think. *Trends in Cognitive Sciences*, *23*(12), 1026–1040. https://doi.org/10.1016/j.tics.2019.09.007

Pillsbury, S. H. (2000). *Judging Evil: Rethinking the Law of Murder and Manslaughter*. NYU Press.

Pinker, S. (2003). *The language instinct: How the mind creates language.* Penguin.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838. https://doi.org/10.1073/pnas.0707192105

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R.

Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil.* (pp. 91–108). American Psychological Association. https://doi.org/10.1037/13091-005

Popp, P., & Gureckis, T. (2020). Ask or tell: Balancing questions and instructions in intuitive teaching. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 1229–1235.

Primoratz, I. (1989). Punishment as Language. *Philosophy*, *64*(248), 187–205.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing* (4.2.1 (2022-06-23)). R Foundation for Statistical Computing. https://www.R-project.org/

Rai, T. S. (2022). Material Benefits Crowd Out Moralistic Punishment. *Psychological Science*, *33*(5), 789–797. https://doi.org/10.1177/09567976211054786

Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences*, *114*(32), 8511–8516. https://doi.org/10.1073/pnas.1705238114

Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences*, *1*, 26.

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, *8*(5), 802–804. https://doi.org/10.1098/rsbl.2012.0470

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*(1), Article 1. https://doi.org/10.1038/ncomms4677

Rangel, A., Camerer, C., & Montague, P. (2008). Rangel A, Camerer C, Montague PR. A framework for studying the neurobiology of value-based decision making. Nat Rev

Neuro 9: 545-556. *Nature Reviews. Neuroscience*, *9*, 545–556. https://doi.org/10.1038/nrn2357

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Roberts, R. M., & Kreuz, R. J. (1994). Why Do People Use Figurative Language? *Psychological Science*, *5*(3), 159–163. https://doi.org/10.1111/j.1467-9280.1994.tb00653.x

Rosen, G. (2004). Skepticism about Moral Responsibility. *Philosophical Perspectives*, *18*(1), 295–313. https://doi.org/10.1111/j.1520-8583.2004.00030.x

Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron*, *70*(6), 1054–1069. https://doi.org/10.1016/j.neuron.2011.05.014

Sarin, A., & Cushman, F. (2022). *One thought too few: Why we punish negligence*. PsyArXiv. https://doi.org/10.31234/osf.io/mj769

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is Organized around Principles of Communicative Inference. *Cognition*, *208*, 104544. https://doi.org/10.1016/j.cognition.2020.104544

Sarin, A., Lagnado, D., & Burgess, P. (2017). The Intention-Outcome Asymmetry Effect: How Incongruent Intentions and Outcomes Influence Judgments of Responsibility and Causality. *Experimental Psychology*, *64*, 124–141. https://doi.org/10.1027/1618-3169/a000359

Scott-Phillips, T. (2014). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Bloomsbury Publishing.

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews Neuroscience*, *8*(4), Article 4. https://doi.org/10.1038/nrn2119

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning From Others: The Consequences of Psychological Reasoning for Human Learning. *Perspectives on Psychological Science*, *7*(4), 341–351. https://doi.org/10.1177/1745691612448481

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. https://doi.org/10.1016/j.cogpsych.2013.12.004

Sher, G. (2009). *Who Knew?: Responsiblity Without Awareness*. Oxford University Press USA.

Sher, G. (2017). Unintentional omissions. In *The ethics and law of omissions* (pp. 3–16).

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *13*, 238–253. https://doi.org/10.1037/h0081183

Shultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *17*, 97–108. https://doi.org/10.1037/h0080138

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of Moral Responsibility and Punishment. *Child Development*, *57*(1), 177–184. https://doi.org/10.2307/1130649

Siegal, M., & Peterson, C. C. (1998). Preschoolers' understanding of lies and innocent and negligent mistakes. *Developmental Psychology*, *34*, 332–341. https://doi.org/10.1037/0012-1649.34.2.332

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211. https://doi.org/10.1016/j.cognition.2017.05.004

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Skillen, A. J. (1980). How to Say Things with Walls. *Philosophy*, *55*(214), 509–523.

Skinner, B. F. (1938). *The Behavior Of Organisms*. Appleton. http://125.18.99.74:8080/jspui/handle/123456789/1377

Smaldino, P., & Richerson, P. (2012). The Origins of Options. *Frontiers in Neuroscience*, *6*. https://www.frontiersin.org/articles/10.3389/fnins.2012.00050

Smith, A. (1869). *The theory of moral sentiments*. Ball & Daldy (Original work published 1759).

Smith, A. (2005). Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics*, *115*(2), 236–271. https://doi.org/10.1086/426957

Smith, A. (2017). Unconscious Omissions, Reasonable Expectations, and Responsibility. In *The ethics and law of omissions* (pp. 36–60).

Smith, P. (1990). Contemplating Failure: The Importance of Unconscious Omission. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *59*(2), 159–176.

Smith, P. (2003). Omission and Responsibility in leal theory. *Legal Theory*, *9*(3), 221–240. https://doi.org/10.1017/S1352325203000107

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Harvard University Press.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*, *173*(5), 1203–1232. https://doi.org/10.1007/s11098-015-0527-9

Sripada, C. (2019). The fallibility paradox. *Social Philosophy and Policy*, *36*(1), 234–248. https://doi.org/10.1017/S0265052519000293

Stephen, J. F. (2014). *A History of the Criminal Law of England*. Cambridge University Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Swannell, S., Martin, G., Scott, J., Gibbons, M., & Gifford, S. (2008). Motivations for Self-Injury in an Adolescent Inpatient Population: Development of a Self-Report Measure. *Australasian Psychiatry : Bulletin of Royal Australian and New Zealand College of Psychiatrists*, *16*, 98–103. https://doi.org/10.1080/10398560701636955

Talbert, M. (2017). *Omission and Attribution Error* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190683450.003.0002

Thorndike, E. L. (1927). The Law of Effect. *The American Journal of Psychology*, *39*(1/4), 212–222. https://doi.org/10.2307/1415413

Thorndike, E. L. (1932). *The fundamentals of learning* (pp. xvii, 638). Teachers College Bureau of Publications. https://doi.org/10.1037/10976-000

Thorndike, E. L. (1998). An Experimental Study of the Associate Processes in .Animals. *American Psychologist*.

Tomasello, M. (2009). *The Cultural Origins of Human Cognition*. Harvard University Press.

Tversky, A., & Kahneman, D. (1973). Availability: Heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tyler, T. R. (2021). *Why People Obey the Law /*.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Villatoro, D., Andrighetto, G., Brandts, J., Nardin, L. G., Sabater-Mir, J., & Conte, R. (2014). The Norm-Signaling Effects of Group Punishment: Combining Agent-Based Simulation and Laboratory Experiments. *Social Science Computer Review*, *32*(3), 334–353. https://doi.org/10.1177/0894439313511396

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and Done? Optimal Decisions From Very Few Samples. *Cognitive Science*, *38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Wagner, A. (1969). *Frustrative nonreward: A variety of punishment* (pp. 157–181).

Walster, E. (1966). Assignment of responsibility for an accident. *Journal of Personality and Social Psychology*, *3*(1), 73–79. https://doi.org/10.1037/h0022733

Wegner, D. M., Giuliano, T., & Hertel, P. T. (1985). Cognitive Interdependence in Close Relationships. In W. Ickes (Ed.), *Compatible and Incompatible Relationships* (pp. 253–276). Springer New York. https://doi.org/10.1007/978-1-4612-5044-9_12

Weiner, B. (1995). *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. Guilford Press.

Weingarten, G. (2010). Fatal Distraction: Forgetting a child in the backseat of a car is a horrifying mistake. Is it a crime? *Children's Voice*, *19*(4), 20–33.

Wellman, H. M. (1992). *The child's theory of mind* (pp. xiii, 358). The MIT Press.

Wellman, H. M. (2017). The Development of Theory of Mind: Historical Reflections. *Child Development Perspectives*, *11*(3), 207–214. https://doi.org/10.1111/cdep.12236

Williams, B., & Bernard, W. (1981). Moral Luck. In *Moral luck: Philosophical papers* (pp. 20–39). Cambridge:Cambridge University Press.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2017). *Mapping the unknown: The spatially correlated multi-armed bandit* [Preprint]. Animal Behavior and Cognition. https://doi.org/10.1101/106286

Xiao, E. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, *77*(1), 321–344. https://doi.org/10.1016/j.geb.2012.10.010

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences*, *106*(28), 11520–11523. https://doi.org/10.1073/pnas.0900636106

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). *Talking with tact: Polite language as a balance between kindness and informativity*.

Young, L., & Saxe, R. (2009). An fMRI Investigation of Spontaneous Mental State Inference for Moral Judgment. *Journal of Cognitive Neuroscience*, *21*(7), 1396–1405. https://doi.org/10.1162/jocn.2009.21137

Zeelenberg, M., & Pieters, R. (2007). A theory of regret regulation 1.0. *Journal of Consumer Psychology*, *17*(1), 3–18.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development*, 2492.

Zhang, Z., Wang, S., Good, M., Hristova, S., Kayser, A. S., & Hsu, M. (2021). Retrieval-constrained valuation: Toward prediction of open-ended decisions. *Proceedings of the National Academy of Sciences*, *118*(20), e2022685118. https://doi.org/10.1073/pnas.2022685118