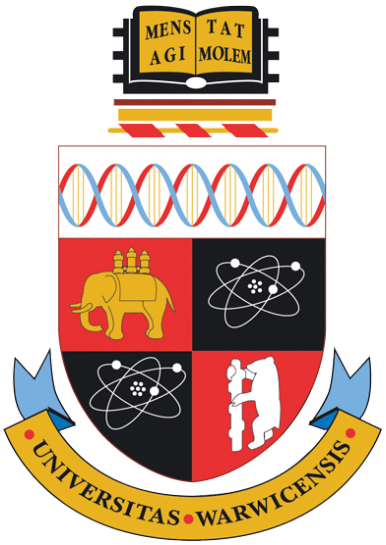
WARWICK
THE UNIVERSITY OF WARWICK



Reacting to Wrongdoers - Victims, Intentionality and Partner Management

Simon Myers

Thesis

Submitted to the University of

Warwick in partial fulfilment of the

requirements for admission to the

degree of

Doctor of Philosophy in Psychology

Department of Psychology

March 2022

Contents

List of Tables and Figures

Tables

Figures

Acknowledgments

First and foremost, I would like to thank my supervisor, Adam Sandborn for always steering me in the right direction. I couldn't have asked for a better supervisor, you were always so generous with your time and always available to advise me when I needed it. You were the opposite of all the horror-stories we seem to be told as prospective PhD students. Thank you! Thanks also to my second supervisor Jesse Preston for always being willing to bluntly tell me when I am being silly and especially for fighting with me over particular pieces of my work that you insisted were worthwhile when I lost confidence in them.

Thanks to the wider faculty at Warwick Psychology, Philosophy and WBS (e.g. Nick Chater, Steve Butterfill, John Michael, Eliot Ludwig and others) and those outside Warwick (e.g. Kurt Gray, Ronnie Janoff-Bullman) for taking time out of their busy days to chat with me and advise me and generally treating me like I was someone worth talking to. The confidence this built provided me the energy to always propel me forward and I wouldn't be the same person without it! Thanks also to Fiery Cushman and his lab members at Harvard for doing the same. I will try and always live up to the esteem you have generously shown me during this time.

Thanks to my family and friends for putting up with me endlessly droning on about moral psychology and testing vignettes and always offering me so much support. Thanks especially to my wonderful girlfriend Abby and my amazing friends Rachael, Liz, Michael, Matt & Tom. I never felt alone during my PhD.

Thanks to everyone here for always being willing to hear my ideas and advise me, even if the ideas were sometimes wacky! I couldn't have done it without you.

Also thanks to these babies, your mews of encouragement never faltered.

Declaration

This thesis is my own work submitted as part of a Doctorate of Philosophy in Psychology from The University of Warwick. It was composed by the author and has not been submitted to any other institution and contains no materials used in other work except those supported by references in the text. The thesis was conducted under the supervision of Dr Adam Sanborn.

Thesis Overview

With regards to the behaviours of our social partners, how do we make judgments of moral relevance? For example, under what conditions do we assign responsibility, level blame, and determine and institute the appropriate responses? How do we determine what constitutes a moral transgression, who the valid victim is and what responses are the most appropriate? These questions lie in the gaps between the conceptual frameworks provided by philosophical ethics, philosophy of mind, and the empirical investigations of the cognitive and social sciences. Leveraging the stereotypical structure of everyday moral judgement (discussed below), this thesis highlights each specific element of that structure and aims to consolidate the most recent research on each element whilst also, across three papers, establishing novel findings for each element.

Kurt Gray and colleagues (2011; 2012; 2014; 2015; 2018) outline what they consider to be the fundamental structure of moral judgement. One [*reacts*] to an [*agent*] who [*intentionally*] [*violates/harms*] a moral [*patient*]. This cognitive template forms part of the Theory of Dyadic Morality (TDM). Although the theory is more extensive, for now, this structure will form the basis for this chapter of this thesis. As written, there are five elements in this structure. 1. Harming/Violating, 2. Moral agents (the principal transgressors), 3. Reactions to those agents, 4. Ascriptions of intentionality, and 5. Moral patients. These five elements are discussed, respectively, as the five sections of the Introduction. Chapters 2, 3 and 4 contain the three papers that form the empirical work for this thesis. Finally Chapter 5 summarises how this novel work fits into the overview established in Chapter 1.

1.1 covers what kind of actions constitute moral harms, how psychologists have disagreed about the centrality of harm over other kinds of violation (e.g., violations that appear to not contain actual physical or emotional harm to a victim) and discusses current pluralist accounts of moral psychology that argue for a range of different kinds of moral violation. 1.2 focuses on judgments that go beyond the act to inferences over the moral agents themselves. Recent work has discovered that a great deal of moral psychology entails character-based rather than act-based judgments and this section discusses how this is important for the understanding of different kinds of moral transgression. 1.3 covers how we react to transgressors including emotional responses, ascriptions of blame and wrongness and how these emotions and judgments drive our different behaviours towards the transgressor. There is a specific focus on the behaviours of Partner Choice (avoidance and ostracization) and Partner Control (punishment). These first sections of Chapter 1 then guide the hypotheses of the first paper, 2.1, *Reacting to Wrongdoers: Harm Leads to Partner Control and Impurity to Partner Choice.*

1.4 covers our current understanding of how people ascribe intentionality, especially how cognitive psychology has highlighted significant deviations from the most prominent normative theories of intentionality ascription. This lays the foundation for the second paper, 2.2, *Does Counterfactual Requirement Explain the Side-Effect Effect?*

Finally, 1.5 covers judgments with regard to the victims of moral transgressions. For example, how do we perceive the minds of moral patients and why does this matter for moral judgement and also whether we are especially sensitive to human moral patients compared to our cousins in the animal kingdom. This establishes the open questions explored in the third paper, 2.3, *Suffering and Dying: How Speciesism Matters for Assessing Extreme Harms.*

Chapter 1. Literature Review

Moral Harms and other Violations

In order to understand what counts as a moral transgression, one must understand what content delineates immoral acts from acts that are merely bad or undesirable (Machery 2008). How can we tell what counts as a violation of a moral norm as opposed to a violation of etiquette or of a prudential norm? Implicitly it is easy, as even by 3 or 4 years of age, young children acquire this capacity and can actively differentiate between conventional and moral transgressions (Nucci and Turiel 1978; Nucci 2001; Tisak and Turiel 1984; Turiel 1977; Turiel 1983; Smetana 2006). However, researchers have struggled to make such delineations systematically. Early research in moral development understood the moral domain around a singular concept of justice (Kohlberg 1973; Haidt 2001; Kohlberg 1971; Kohlberg and Hersh 1977). However, this proved to be insufficient on its own, both because it was vaguely defined and also because it left out other important aspects like care in interpersonal relationships (Gilligan 1994) and the cultural and religious norms that surround moral judgement but which are not easily understood in terms of justice (Damon 1999; Graham et al. 2018; Graham et al. 2011; Haidt and Joseph 2008).

To expand on what counts as a moral consideration, Turiel (1983) identified that the moral domain includes, at least, the concept of harm, fair distributions of resources and the protection of the rights of individuals. In addition, he showed that moral violations, unlike conventional violations, are regarded as wrong independent of authority and are seen to be more serious, and more universal. Although these findings show that people identify the moral domain as a distinct entity, apart from other social reasoning, they do not fully distinguish the necessary content that people use to identify an act as immoral.

More recently, two popular accounts have been proposed to address this, the Theory of Dyadic Morality (TDM) and Moral Foundations Theory (MFT). TDM argues that morality is grounded exclusively by the presence of intentional harm and nothing more (Schein and Gray 2018; Gray et al. 2014; Schein and Gray 2015; Hester et al. 2020). When there is a vulnerable victim who is acted upon by an intentional agent, whereby that victim appears to suffer in some way because of the action, people perceive harm and automatically judge the action to be morally wrong, even if it is entirely ambiguous as to what the act was (Hester et al. 2020). This is to say that it is not the content that matters but rather the structure, where two agents are causally connected by way of one causing intentional or avoidable damage (physical or psychological) to the other. More recently, proponents of TDM expand the idea of a victim to consist of depersonalised entities such as the “social order” or spiritual entities such as souls ((Schein and Gray 2018; Gray et al. 2022). In addition, the extent to which one judges an act as immoral is driven by how much harm one perceives (Schein and Gray 2018). For some cultural norms it is often difficult to identify a harmed party, for example in certain cases of moralising sexuality, or particularly weird or disgusting acts that people identify as morally prohibited. Indeed some of these acts are “objectively” harmless, however, Schein et al. (2016) demonstrate that even in these cases, the level to which people find them immoral is mediated by intuitive and automatic perceptions of harm. Also, the perceived immorality of acts that have caused no objective harm is substantially predicted by perceptions that such acts risk causing harm, in general, and drawing attention to these risks increases ascriptions of immorality (Stanley et al. 2019). From obvious harmful violations such as murder or assault to less clear violations like dishonour, pornography or a lack of patriotism, TDM argues that the extent to which we recogise these things as immoral is fundamentally grounded in perceptions of harm.

TDM is in direct opposition to Moral Foundations Theory (Graham et al. 2018; Graham et al. 2011; Haidt and Joseph 2008; Koleva et al. 2012; Graham et al. 2009). MFT argues morality is encompassed by five or potentially more separate and distinguishable categories, of which harm is but one, the others being fairness, loyalty, authority and purity, although this is not meant to be an exhaustive list. By leveraging people's apparent differing sensitivity towards specific categories, Moral Foundations theorists have shed light on many different areas of moral psychology, such as persuasion and moral reframing (Feinberg and Willer 2019), environmentalism (Feinberg and Willer 2013), vaccine hesitancy (Amin et al. 2017), and our taste for different culture war narratives (Koleva et al. 2012). Moral Foundations theorists argue that the historical focus on harm and fairness comes from the tendency of social psychologists to be politically left-leaning and therefore less sensitive to the other foundations (Graham et al. 2009; Kivikangas et al. 2020). Graham et al. (2018) also argue that most participant pools are from WEIRD cultures (Western, educated, industrialised, rich and democratic [Gilligan 1994; Henrich et al. 2010]) failing to capture the full spectrum of moral concern that exists across cultures.

Work conducted as part of this thesis (see Chapter 2) has explored a novel intermediary theory which attempts to adjudicate between the harm-centric TDM and the more pluralistic MFT. We posit, in agreement with TDM, that morality revolves around conceptions of harm. However, we argue in agreement with MFT, that ostensibly harmless wrongs (in particular violations of moral purity) are indeed a separate category in that they stem from judgments and behaviours that are intended to avoid *future* harm from undesirable social partners, because of the character-based signals inferred from actions such as purity violations. We discuss character-based judgments in the next section (1.2).

The most widely discussed and debated foundation is that of moral purity (e.g. Gray et al. 2021; Chapman and Anderson 2013; van Leeuwen et al. 2012; Sabo and Giner-Sorolla 2017; Wagemans et al. 2017; Giner-Sorolla and Chapman 2017; Russell and Giner-Sorolla 2011; Dungan et al. 2017; Tepe and Aydinli-Karakulak 2019). Purity violations have been argued to be rooted in feelings of disgust but the connection is more complicated than that, especially with regards to religious-based violations where where several different emotional reactions are involved (Royzman and Kurzban 2011; Royzman et al. 2014; Kollareth et al. 2021; Kollareth and Russell 2019). There is little consensus regarding both the definition of moral purity and its operationalization (Gray et al. 2021). Indeed, the way purity is measured is often inconsistent with the way it is defined (Gray et al. 2021). Purity violations include acts related to sex, e.g., loving incest (Haidt 2001), necrophilia, bestiality or an individual having sex with a frozen chicken before cooking it for dinner (Clifford et al. 2015); food contamination (Haidt 2001; Graham et al. 2018); protecting individuals from pathogens; protecting sacred objects, rituals and concepts from spiritual deflilment (Preston and Ritter 2012); and thinking profane or blasphamous thoughts (Schein et al. 2016). For a review of the heterogeneous definitions and opperationalistions of moral purity see (Gray et al. 2021). Evolutionary views see moral disgust as an exaptation to control alien ideas, free riders, or repeat moral violators as well as literal pathogens (Tybur et al. 2013; Chapman and Anderson 2013). Most importantly, purity is defined in direct opposition to harm (Haidt 2012; Haidt and Hersh 2001; Graham et al. 2009; Björklund et al. 2000) where acts are assumed to not include an identifiable victim or harmed party (Chakroff, 2015; Gutierrez & Giner-Sorolla, 2007), and is most closely linked to the emotion of disgust (Inbar et al. 2009; Graham et al. 2013; Haidt et al. 1993; Horberg et al. 2009). In contrast, harm is more related to the emotion of anger (Horberg et al. 2009; Inbar et al. 2009; Giner-Sorolla and Chapman 2017). It is argued that this is because moral purity evolved as a "behavioural immune system" in order to socialise pathogen avoidance (van Leeuwen et al. 2012; Schaller and Park 2011; Thornhill and Fincher 2014; Tybur et al. 2016; Fincher and Thornhill 2012). However, note that this struggles to account for divinity concerns (Piazza et al. 2018; Preston and Ritter 2012; Piazza et al. 2018; Ritter et al. 2016) or

concerns over order and xenophobia (although it may be argued that xenophobia does have some ties with pathogen avoidance) (Graham et al. 2018; Graham et al. 2011; Janoff-Bulman and Carnes 2013; Haidt 2007). Note also that disgust is not solely implicated in moral purity but rather occurs with particular character-based judgments, even when they are based in harm rather than purity (Giner-Sorolla and Chapman 2017).

Character-Based Judgements of the Moral Agent

Apart from focusing on the content of blameworthy acts that people identify as moral violations, another important aspect of moral judgement centres on the agent's moral character. Understanding these character-based or personological judgments is crucial because recent research has shown they have important and often overlooked implications for moral judgements (Uhlmann et al. 2013; Uhlmann et al. 2015; Pizarro et al. 2003; Everett et al. 2016; Pizarro and Tannenbaum 2012; 2011). Character-based judgments rely on the fact that the strength of our moral condemnation is often driven by the extent to which we learn undesirable features of an agent's moral character rather than the level of consequential harm one perceives. This view takes its origins from Aristotelian virtue ethics as well as Hume's views on moral character (Uhlmann et al. 2015; Sripada 2010). When interacting in the social world we are motivated to learn as much as much as we can about our social partners and so will be focussed on the character-based information that certain acts convey. Personological research has sought to empirically distinguish between these kinds of judgments and the more traditionally discussed act-based judgments. Purity violations are unusual and so are understood to be statistically rare making them particularly important in diagnosing moral character (Fiske 1980; Ditto and Jemmott 1989; Walker et al. 2020; Wagemans et al. 2017; Gray and Keeney 2015).The fact that purity violations are argued to be merely "weird" (Gray and Keeney 2015) can be a legitimate feature for diagnosing moral character (Uhlmann et al. 2015). Indeed, purity violations are seen to be particularly diagnostic because people judge them to be less driven by situational factors or context (Cushman 2008; Chakroff et al. 2013; Chakroff and Young 2015; Russell and Giner-Sorolla 2011).

Consider a corporate executive who extravagantly purchases luxury items like a giant yacht or a set of diamond encrusted shower curtains for tens of thousands of dollars. Such acts may elicit public outrage without necessarily causing a great degree of harm in the world (Landy and Uhlmann 2018; 2011). Similarly, Tannenbaum et al., (2011) asked participants to consider a CEO using their large bonus to purchase a personalised marble table for their office featuring an engraving of their own face. For these kinds of acts, negative judgments are not formed over consequential harms but rather centre on the moral character of the person making the request. Or instead consider a bigot who directs his ire towards a few members of a racial minority rather an a general misanthrope who directs his ire to a larger amount of people without discrimination. Despite harming fewer people, participants made harsher judgments of the bigot, specifically because they were able to learn more about his moral character (Uhlmann et al. 2014). Even those perceived to have done the right thing (from a consequentialist perspective) can be seen as having deficient moral character and this is true even when the act itself is judged to be praiseworthy (Uhlmann et al. 2013). Chapter 2 explores the different character-based signals inferred from different kinds of moral violation and their utility for managing our social partners.

Reacting to Moral Transgressions

How are our reactions and behaviours influenced by the kinds of moral judgement examined in the previous sections? As discussed, purity violations are more often met with disgust and harm violations are more often met with anger. These emotions help regulate two opposing behavioural systems, the

approach-based Behavioural Activation System (BAS) and the avoidance-based Behavioural Inhibition System (BIS) (Carver and White 1994; Kim and Lee 2011; Corr 2002). The BAS regulates behaviours that are directed towards some desired goal where the relevant behaviours are activated in order to obtain that goal (Carver and Harmon-Jones 2009). For example, it can cause one to move towards a particular desired food item. The BIS inhibits behaviours that are related to aversive outcomes in order to move away from unpleasant things that may cause such outcomes (Shook et al. 2019). For example, fear can be activated to prevent approaching a predator, and disgust can be activated in order to avoid consuming tainted or rotting food. Notably, the BIS and BAS have previously been shown to be important in regulating moral behaviour (Janoff-Bulman et al. 2009; Janoff-Bulman and Carnes 2013; Sheikh and Janoff-Bulman 2010). For example, the BAS can regulate guilt in order to motivate one to approach a social partner to make amends for a transgression and the BIS can activate shame causing one to avoid and shy away from their social partners (Sheikh and Janoff-Bulman 2010). Proscriptive morality (morally bad actions one should not do) is inhibited by the BIS avoid system where they are seen to be strict and concrete, whereas prescriptive morality (morally good actions one ought to do) are activated by BAS, but they are done so more weakly and are seen to be abstract and discretionary (Janoff-Bulman et al. 2009). It is important to note that although disgust and anger are discussed here as separate emotions, they can co-occur and be entangled with each other making such separation conceptually and methodologically difficult to tease apart (Russell and Giner-Sorolla 2011; Gutierrez and Giner-Sorolla 2007; Chapman and Anderson 2013).

In Chapter 2, the paper explores the possibility that a similar relationship may be found when making judgments of third parties, positing that when witnessing or learning about a moral violation, anger would cause one to approach the wrong-doers whereas a disgust response would result in avoiding the wrong-doers. For example, the BAS can regulate anger so one can aggressively react to a transgressor, seeking retribution, and the BIS can regulate moral disgust so one avoids social partners for whom negative character based judgements have been made. This contrast of approaching versus avoiding social partners is characteristic of the behavioural strategies of Partner Choice and Partner Control which reinforce cooperation between social partners (Bull and Rice 1991; Campennì and Schino 2014; Noë 2006; Martin et al. 2020; Barclay and Raihani 2016; Martin et al. 2019).

When engaging in Partner Control, agents directly reward good behaviour or cooperation through direct reciprocity or punish those who engage in undesirable behaviour to prevent its recurrence. Partner Choice is less direct and based on a more holistic account of past behaviours. Here, social partners select agents who appear to be the most willing to cooperate and avoid those who are not, leading to competition for selection in the biological market (Barclay 2016; Martin et al. 2019). Both Partner Control and Partner Choice behaviours can be driven by the inferred intentions of the agent, but Partner Control can be particularly reactive to even unintended outcomes (Barclay and Raihani 2016; Martin et al. 2019; Martin and Cushman 2015). In other words, when engaging in Partner Control the focus is more on consequential outcomes but when engaging in Partner Choice the focus is more on diagnosing the moral character of potential social partners.

Ascribing Intentions

If we wish to learn about the character of a moral agent by their actions, how can we tell when their actions were actually intentional? For that matter, what does it mean to say that an outcome was brought about intentionally? To ascribe intentionality, one makes inferences over the state of mind of the actor, reasoning about whether the agent's state of mind was in some sense directed towards the outcome in question. Predominantly these inferences consider the agent's foreknowledge, awareness, skills, desires and beliefs (Malle and Knobe 1997). Purportedly, it would at least require that the agent desired some

outcome and believed that their action would obtain it. However, probing individuals' ascriptions of intentionality reveals a more complex picture that is difficult to explain with simple definitions of folk intentionality (Machery 2008; Cushman and Mele 2008; Cova 2016). In a now famous experiment, Joshua Knobe (2003), asked participants to consider one of two situations where a CEO implements a new policy that will make lots of money but, as a side-effect, will affect the environment. Suppose also that the CEO is indifferent to environmental concerns. When the policy is said to harm the environment, 82% of participants judge the harm to be intentional. Remarkably however, if the policy instead helps the environment (everything else is kept the same) only 23% ascribe intentionality. This effect is called the Knobe effect or Side-Effect Effect (SEE). It appears as if inferences are not simply made about the actor's state of mind but rather people may be sensitive to the valence of the outcomes (Knobe 2003) or normative considerations (Hindriks 2014; Knobe 2010).

In a similar experiment, Knobe (2003) asks participants to consider someone who wins a rifle shooting contest by hitting the bull's eye. If he does so skillfully, 79% of people say it was done intentionally, but if he only manages to hit his target through sheer luck, only 29% ascribe intentionality. This is contrasted with someone who wishes to inherit his aunt's money by shooting and killing her. Here, he either kills her through skillful marksmanship and 95% of people agree it was intentional, but unlike the shooting competition, when he hits his target through sheer luck, the substantial majority (76%) still attribute intentionality. This is called the Skill Effect (SE).

These strange asymmetries were replicated cross-culturally (Burra and Knobe 2006; Cova and Naar 2012) and with young children (Leslie et al. 2006; Pellizzoni et al. 2009). In addition the asymmetry exists when ascribing intentionality to groups (Michael and Szigeti 2019). Similar effects appear to also exist for judgments of knowledge (Paprzycka-Hausman 2018; Dalbauer and Hergovich 2013) and freedom (Phillips and Knobe 2018). The many competing explanations demonstrate a striking lack of consensus for understanding folk intentionality (e.g.Laurent et al. 2019; Nakamura 2018; Quillien and German 2021; Dalbauer and Hergovich 2013; Laurent et al. 2020; Cova 2016; Lin et al. 2019; Hindriks et al. 2016).

Are accurate judgments being biassed by considerations that should be orthogonal to intentionality? Early accounts argued that these asymmetries reflect biases from being motivated to make judgments about blameworthiness (Alicke 2008; Nadelhoffer 2004) or responsibility (Wright and Bengson 2009) or are biassed simply by negative affect (Nadelhoffer 2004). However, these accounts have fallen out of favour because they fail to predict responses to slightly different vignettes or fail to predict responses when the source of bias has been removed (for a review see [Cova 2016]). Another prominent explanation is to suppose intentionality captures multiple concepts and people infer what concept is being asked about by the context of the vignette. This is called the Interpretive Diversity Hypothesis (IDH), (Laurent et al. 2020; Laurent et al. 2019; Nichols and Ulatowski 2007; Cova 2016; Cushman and Mele 2008). The IDH argues that in some cases people consider whether an act was done "intentionally" but in other cases, for example when someone does something morally wrong, one believes they are being asked about whether the act was "done with foreknowledge" (Nichols and Ulatowski 2007; Laurent et al. 2020; Laurent et al. 2019). Cushman and Mele (2008) propose that there are "two and a half folk concepts" of intentionality where one examines either desires, beliefs or some interaction between them depending on the context.

More recent work has highlighted the importance of counterfactual reasoning in these kinds of judgement (Kim et al. 2018; Halpern and Kleiman-Weiner 2018; Kleiman-Weiner et al. 2015; Quillien and German 2021; Phillips et al. 2015). Kleiman-Weiner et al. (2015) propose that when judging the intentionality of a side-effect, one reasons about whether the side-effect was counterfactually required for the desired outcome. This means that one thinks about slightly different states of affairs where the

side-effect would not occur and checks whether the desired outcome would also fail to occur. If the side-effect and desired outcome are linked in this way, then that means that the side-effect was counterfactually required to bring about the desired outcome. In these cases people would judge the side-effects to be intentional.

To illustrate this, consider the classic trolley dilemma where someone pulls a lever to divert a trolley onto an adjacent track so that the five people who would have been run over by the trolley would be saved. Unfortunately, a bystander is on the adjacent track and so, as a side-effect, they are killed instead. If one imagines worlds where the bystander was not on the adjacent track, pulling the lever would still save five people so the bystanders death was not counterfactually required. Instead, suppose the scenario is slightly different such that the adjacent track connects back to the first track. If the lever is pulled, the trolley switches to the adjacent track and hits the bystander causing the trolley to stop and the five people are saved. However, if there is no bystander then the trolley won't stop and will loop back onto the main track killing the five people. In this adjusted version of the trolley dilemma, killing the bystander is counterfactually required to save the five. Importantantly, participants judge counterfactually required side-effects in trolley dilemmas as indeed more intentional (Kleiman-Weiner et al. 2015). The paper attached to Chapter 3 tests this model on the classic SEE cases.

An alternative account from Quillien and German (2021) also highlights the importance of counterfactual reasoning but does so by borrowing from a particular philosophical, normative theory of intentionality (Davidson 1980; Davidson 1963) and updating it with recent developments in causal cognition (Quillien 2020; Icard et al. 2017; Morris et al. 2019; Willemsen and Kirfel 2019; Lagnado et al. 2013). The repurposed philosophical account is a causalist account, stating that outcomes are intentional when the agent's attitudes towards those outcomes (desires, beliefs and expected utilities) *cause* the outcome. Causal selection in human cognition also relies on counterfactual reasoning (Quillien 2020; Icard et al. 2017; Morris et al. 2019; Willemsen and Kirfel 2019; Lagnado et al. 2013). Chapter 3 discusses how this model may be able to account for the data from the Kleinman-Weiner model and also the new data from the empirical work conducted for this thesis.

Moral Patiency

Now we turn to the final feature of the structure of everyday moral reasoning, that of the moral patient, who or what is the victim of a moral transgression and why this matters for moral judgments. When given a clear description of a specific person in jeopardy or suffering, we are far more likely to care and be willing to expend resources than if that individual is presented merely as a statistic (Jenni and Loewenstein 1997; Schelling 1968; Small and Loewenstein 2003; Small and Loewenstein 2005). The more we know about the victim the more we care (Schelling 1968; Jenni and Loewenstein 1997). Descriptions of a victim can elicit empathy and compassion and motivate us to a have moral concern for them (Coke et al. 1978; Toi and Batson 1982; Batson et al. 2002) but when given descriptions of too many victims we become numbed to their plight (Fetherstonhaugh et al. 1997).

TDM argues that the cognitive template for moral judgement requires one perceives two complementary minds (or depersonalised entities such as the "social order" that can be perceived to be harmed), that of the moral agent and the moral patient (Gray et al. 2012; Schein and Gray 2018; Gray et al. 2022). Before inferring whether a patient is harmed or is suffering, mind perception must occur (Epley and Waytz 2009). We perceive groups to have "less mind" than individuals (Cooley et al. 2017; Knobe and Prinz 2008) which may help explain why we appear to care more about specific individuals, precisely because we are more able to perceive their minds.

Mind Perception has been argued to be “the essence of moral psychology” (Gray et al. 2012), but how does the psychology of mind perception work? Gray and colleagues (2007) argue that there are two dimensions to the kinds of minds people automatically attribute, which predict the kinds of moral rights we believe those agents have. The first dimension is experience (or patiency) which is the ability to have internal experiences like hunger, pleasure or pain. The second dimension is agency, which relates to the ability to think, plan and communicate and also the ability to control one’s actions. Importantly, different animal species are seen to vary greatly across these dimensions as well as in comparison to humans (Gray et al. 2007).

While cute animals are seen almost entirely as moral patients (high experience - low agency), humans generally have both high agency and patiency and can be easily imagined both as victims or as blameworthy perpetrators. Similarly, Kara Weisman and colleagues (2017; 2021) argue that there are three dimensions to our mental capacities: body, heart and mind. Each of these relate to aspects of both experience and agency (Weisman et al. 2017; Weisman et al. 2021), and they have also been conceptualised as bodily sensation, cognition and emotion (Malle 2021). People even ascribe minds based on the speed of movement the agent appears to have (Morewedge et al. 2007). Targets that move at speeds similar to human movement were perceived to be more likely to have mental states. People also show an inverse relationship in perceptions of agency and patiency, known as Moral Typecasting (Gray and Wegner 2009). People who are perceived to be moral patients (high experience) are less likely to be seen as being capable of being perpetrators, whereas we are less able to see those with high agency as victims, although see Arico (2012) for criticism of this work.

What kinds of judgement may be sensitive to who or what the moral patient is? In Chapter 4, we consider cases of intentional killing compared to cases of torture (causing a moral patient to suffer but not die). Philosophers have found it to be a far from trivial task to isolate judgments about killing when they are removed from considerations about suffering. In the original formulations of Utilitarianism, the concept of utility was centred on positive and negative mental states like happiness and pain (Bentham 1789; Mill 1887; Sidgwick 1981). This is referred to as Hedonistic Utilitarianism. On this view, it is difficult to account for why we care so much more about acts of killing (Henson 1971). For example, imagine that a vagrant with no family or friends is painlessly murdered in their sleep. Because this entails no explicit suffering, it is difficult for theories that centre on mental states to account for why we feel this act is particularly morally wrong (Carson 1983). Indeed it has been argued that the only moral issue with painlessly wiping out all life is simply the prospect of failure, as failing could lead to immense suffering (Pearce 2005). Besides explicitly specifying extra deontological constraints against killing, the most common response to this comes from Preference Utilitarianism, where the concept of utility focuses instead on the fulfilment and frustration of the agent’s preferences (Singer 2011; Hare 1981; Singer 1995). Here, what is important is the vagrant's preference to continue living rather than their feelings of happiness or suffering. The fact that humans have much larger and more complex preferences compared to animals is integral to why we believe human life to be of particular importance (Singer 2011). Singer (2011) also highlighted the moral difference between killing a being who experiences pain but is not able to meaningfully reflect about themselves, compared to killing a more intelligent being. This contrasts with Jeremey Betham’s famous declaration, “the question is not, can they reason, nor can they talk, but, can they suffer?” (Bentham 1789).

The dimensions of mind perception may be able to explain why we care more for the welfare of humans than other animals. If we require that rights only be possessed by what we consider to be “rational beings” (Kant and Schneewind 2002; Korsgaard 2004), then it is possible that the lower agency we attribute to animals (given that we perceive them to be less rational) reduces our motivation to protect them over

humans. Indeed, it has been argued that animals, unlike humans, do not have the deontological right to not be sacrificed for the greater good (Thompson 1990). In addition, the lower experience we attribute to animals may drive us to care less about their suffering simply because we perceive them to have a lower capacity to suffer than humans. The idea that humans possess unique features salient to moral concern is well studied in philosophy of ethics (e.g. Kagan 2016; Grau 2016).

However, there may be more fundamental reasons why we privilege the moral status of humans. Afterall, people show anthropocentric biases, perceiving lots of natural phenomena as being designed for the benefit of humans over other animals (Preston and Shin 2021). The view that moral worth is determined, prejudicially, only on the basis of species membership is called Speciesism (Horta 2010; Caviola et al. 2019; Caviola et al. 2020; Fjellstrom 2002), likened to racial prejudice or sexism (Fjellstrom 2002; Caviola et al. 2019). Despite the fact that mind perception goes some of the way to explain our preference for the protection of human welfare, Speciesism better predicts moral worth judgments over and above this (Caviola et al. 2019). In Chapter 4 we explore how the relative contributions of Speciesism and mind perception differ across different animal and human moral patients when making judgments about killing compared to judgments about torture.

The following three chapters are the three research papers which make up the empirical work of this thesis.

Chapter 2. Reacting to Wrong-Doers: Harm Leads to Partner Control and Impurity to Partner Choice

Abstract

We explore two types of behavioural strategies that individuals use to manage their social partners in response to undesirable behaviour. These strategies are Partner Control - rewarding and punishing behaviour, and Partner Choice - selecting preferred social partners and avoiding the others. Past work has shown that Partner Control and Partner Choice are functionally distinct but this research has not investigated whether people use different strategies to address different types of undesirable behaviour. In particular, we focus on violations of Harm and Purity (from Moral Foundations Theory). Using a selection of Harm and Purity vignettes, across two preregistered studies we test how different violations result in a preference for one Partner Management strategy over the other, based on the signals people infer from these actions, while exploring the role of theoretically relevant demographic measures (political orientation, Relational Mobility, urban vs rural residence). We find that Harm violations are more likely to lead to Partner Control and Purity violations are more likely to lead to Partner Choice, particularly when impurity gives strong character-based signals. By appealing to these contrasting strategies, we propose a functional account of why people differ in their sensitivity to moral purity, especially victimless wrongs. We theorised that these individual differences may be explained by the tightness of social groups and how much opportunity they provide for Partner Choice, although our preliminary evidence does not yet support this.

Introduction

Suppose you learn that one of your friends has broken into someone's house in order to steal their property. Or suppose that you discover a different friend frantically searching through the trash, and when asked to explain why, he gleefully says it is to find a stranger’s discarded underwear. Assuming that you find each of these actions to be in some way morally questionable, how would you expect to react to each friend? One reaction to a moral violation is to seek retribution by punishing the perpetrator, another is to avoid them, no longer including them in the social group and instead seeking out others for joint tasks, partnerships and social group membership. These two strategies for social partner management have been formalised as Partner Control and Partner Choice respectively (Bull and Rice 1991; Campennì and Schino 2014; Noë 2006).

Partner Control and Partner Choice are mechanisms to bolster cooperation between social partners. Partner Control is characteristically reactive; an agent modulates a partner's behaviour by directly rewarding cooperation through reciprocity or punishing undesirable behaviour to prevent its recurrence. Partner Choice, on the other hand, involves selecting or deselecting partners based on evaluations of their past behaviour; agents signal their willingness to cooperate in an attempt to outcompete each other for selection in a biological market so they can benefit from increased social partnership and joint cooperation (Barclay 2016; Martin et al. 2019). These two mechanisms are functionally distinct. While engaging in Partner Choice people are focussed on the intentions of the agent, whereas when people are engaging in Partner Control they are more likely to react to outcomes even when they are not necessarily intended (Barclay and Raihani 2016; Martin et al. 2019; Martin and Cushman 2015). In other words, when engaging in Partner Control the focus is on consequential outcomes but when engaging in Partner Choice the focus is on the character and intentions of the social partner. These strategies may not be entirely

distinct, for example punishment may be followed by exclusion and in turn the social partner may be offered a chance to return. In addition, exclusion may be used as a threat as part of Partner Control.

Despite the importance of these two strategies, past work has not investigated whether the strategy being employed depends on the type of moral violation it is in response to. Examples of different kinds of moral violation are the aforementioned burglary versus the commandeering of the stranger's discarded underwear presented above. Moral Foundations Theory identifies five separable categories of morality, and the two which are contrasted the most are the foundation of harm (which corresponds to the former violation) and purity (which corresponds to the latter) (Graham et al., 2011; Haidt & Joseph, 2008). Harm involves a moral agent intentionally causing distress or injury to a moral patient (Gray et al., 2014), whereas purity violations often contain so-called "harmless wrongs" (Gray et al., 2014) or "victimless wrongs" (Chakroff, 2015; Gutierrez & Giner-Sorolla, 2007) where a moral agent does something disgusting, weird, or counter-normative. MFT has provided a great deal of insight into our moral psychology in areas such as persuasion through moral reframing (Feinberg and Willer 2019), attitudes towards environmentalism (Feinberg and Willer 2013), vaccine hesitancy (Amin et al. 2017), and culture war narratives (Koleva et al. 2012).

Why might we expect people to react differently to the perpetrators of different kinds of moral violation? One possibility is that people choose either Partner Control or Partner Choice based on how morally wrong they believe a violation to be. For example, it may be the case that purity violations, on average, are seen as less morally wrong and so on average they are met with a different strategy than harm violations. If this is the case then specific purity violations that are seen to be especially morally wrong may be treated the same way as harm violations. It is not clear, however, whether Partner Control or Partner Choice is the stronger reaction and therefore it is hard to ascertain what the most appropriate response is to especially wrong violations. It is also possible that moral wrongness is not the most important measure in discerning between Partner Management strategies.

Another candidate which may drive preferences for different Partner Management strategies is that different violations may elicit different emotional reactions and these in turn may drive either Partner Choice or Partner Control. Notably, the foundations of Purity and Harm have been linked to specific emotions that are important for behavioural regulation. Purity seems to be more closely associated with the emotion of disgust, whereas anger more often results from Harm (Horberg et al. 2009; Inbar et al. 2009; Giner-Sorolla and Chapman 2017). In addition, anger is associated with direct aggression, especially when one imagines themselves as the victim, and disgust is associated with less costly indirect aggression, especially when one imagines a third-party as the victim (Molho et al. 2017). Anger is an approach-based emotion, linked to the Behavioural Activation System (BAS), (Carver and Harmon-Jones 2009), where the agent is more likely to move towards the target. Disgust, on the other hand, is more closely related to the Behavioural Inhibition System (BIS) where actions that would otherwise lead towards the target are suppressed (Shook et al. 2019). The BIS and BAS systems have been implicated in the regulation of our own moral behaviours, where morally bad actions are inhibited and morally good actions are activated (Janoff-Bulman and Carnes 2013; Sheikh and Janoff-Bulman 2010; Janoff-Bulman et al. 2009). A similar relationship may be found when making judgments of third parties: when witnessing or learning about a moral violation, anger would cause one to approach the wrong-doers, in order to engage in Partner Control, whereas a disgust response would result in avoiding the wrong-doers, a form of Partner Choice. It is important to note that although disgust and anger are discussed here as separate emotions, they can co-occur and be entangled with each other making such separation conceptually and methodologically

difficult to tease apart (Russell and Giner-Sorolla 2011; Gutierrez and Giner-Sorolla 2007; Chapman and Anderson 2013).

Importantly, moral disgust is not simply activated by purity violations but can be a more general reaction to negative information about moral character (Giner-Sorolla and Chapman 2017). Although purity violations are often diagnostic of character, disgust is also generated when the harm violator's *desire* to cause harm is made particularly salient. An explicit desire to do wrong without good reason can signal information about the moral character of the perpetrator in the same way purity violations in general can. In other words, consequential harms will generally generate anger whereas more explicit signals of moral character can generate disgust. This distinction between act-based and character-based judgments can provide a reasonable basis for the different emotional reactions.

Moral judgments that centre on the agent's moral character (apart from the moral wrongness and emotional response) are called *personological* judgments and have been identified as an important and often overlooked factor in moral judgement (Uhlmann et al. 2013; Uhlmann et al. 2015). Tannenbaum et al., (2011) asked participants to consider a CEO requesting to have their large bonus be put towards a personalised marble table for their office featuring an engraving of their own face. This is contrasted with a CEO who only gets the monetary bonus. Here, greater negative judgments of the first CEO compared to the second are not formed over a blameworthy act or a consequential harm but rather centre on the moral character of the person making the request. Purity violations, in particular, are seen to be more revealing about the perpetrator than harm violations and less driven by situational factors or the context of the action (Cushman 2008; Chakroff et al. 2013; Chakroff and Young 2015). In addition, people find harm violations to be more easily justifiable by contextual factors (Russell and Giner-Sorolla 2011). Even simply imagining purity violations is judged to be morally worse than imagining harm violations (Sabo and Giner-Sorolla 2017). Judgments of accidental harms are seen as less morally wrong compared to accidental purity violations (Young and Saxe 2011) because it is argued, a purity violation still leaves the violator contaminated (Dungan et al. 2017; Appiah 2007). Relational context also has differing effects on judgments of purity compared to harm violations. For example, when a perpetrator acts alone and violates themselves, self-inflicted impurity results in stronger negative moral judgments compared to self-inflicted harm (Dungan et al. 2017; Chakroff et al. 2013).

It is possible then that the decision to engage in a specific Partner Management strategy could be driven by the strength of character-based judgments. But why might this be? Chakroff and colleagues (2015), examined how moral violations are predictive of the agent's future behaviour. Agents who performed impure acts were judged as more likely to perform impure acts in the future and, notably, they were also judged as more likely to be harmful in the future. Harmful acts, on the other hand, were only deemed to be predictive of future harm but not future impurity. It is possible then that character-based judgments that specifically signal the potential for future harm (when there is no current harm) will be decisive in whether people engage in Partner Choice compared to Partner Control.

Another possibility is that character-based judgments are useful for determining whether a social partner will be predictable, which is useful for future social collaboration (Walker et al. 2020). This may centre on how easy it is to understand the motivations of a moral violator. For example, a moral violation driven by greed or jealousy may be more easily understandable than purity violations that are more associated with actions that are perceived to be disgusting.

To recap, some of the potential drivers of the differing Partner Management strategies that have been discussed so far are judgments of moral wrongness and character diagnosticity, emotional reactions,

judgments regarding the motivations of the violator, perceptions of their predictability and concerns over future harm (when there is no present harm). It is likely that these elements are not easily separable. For example, character-based judgments may drive perceptions of predictability but these may also stem from an inability to understand or identify with the motivations which led up to the act. In addition, many of these may correlate closely with each other, for example, highly morally wrong acts may be seen to be especially revealing of moral character. Despite this difficulty, it is important to establish what combination of these factors predict the chosen Partner Management strategy and also to understand how each of these elements may drive the strength of the given reaction.

Aside from situational factors, we might expect individual differences to play a role in people's partner management reactions. There are individual differences between people's sensitivity towards purity violations, namely, those who identify as political liberals are less sensitive to violations of moral purity than those who identify as political conservatives (Graham et al. 2009; Kivikangas et al. 2020). Liberals appear to be more concerned with wrongs where there is some objective harm to a victim, and they are less sensitive than conservatives to wrongs for which there is no easily identifiable victim (which is most often the case with purity violations) (Feinberg & Willer, 2019; Graham et al., 2009; Hatemi et al., 2019; Kivikangas et al., 2020). Note however that these political differences in the sensitivity to different foundations is more fine grained such that under certain circumstances liberals will endorse moral views based on sanctity whereas conservatives may endorse views on the same topic based on fairness (Frimer et al. 2017). In addition, liberals will morally condemn certain "harmless" wrongs more relevant to them (Frimer et al. 2015).

The tendency to engage in Partner Choice over Partner Control might also be sensitive to the extent to which people have greater or fewer options for social partners or whether avoiding or ostracising social partners is particularly costly. Relational Mobility is a measure of how much freedom and opportunity a society affords people to form new social relationships (Thomson et al. 2018). In societies with low Relational Mobility (e.g. rural farming areas) it is particularly difficult to dissolve one's current interpersonal relationships whereas in high Relational Mobility societies (e.g. urban cities) there is greater freedom to select social partners and avoid others (Thomson et al. 2018). Because of this we also assessed the Urban-Rural divide between our participants (Nemerever and Rogers 2021). Conservatives, compared to Liberals, are more likely to live in rural areas where communities and social groups are tighter (Maxwell, 2019; Waytz et al., 2019). Rural areas may be more isolated and/or sparsely populated and so are likely to provide fewer opportunities for Partner Choice so a greater sensitivity to signals of moral character that would indicate the potential for future harm may be particularly beneficial when it is harder to avoid problematic social partners. Conversely, signals of moral character may be particularly beneficial if one has lots of opportunities for Partner Choice because it provides a basis to distinguish options.

The Current Studies

We hypothesise that Purity violations will more often lead to Partner Choice and Harm violations will more often lead to Partner Control, perhaps because stronger character-based judgments are made for Purity violators, as these kinds of acts give a strong signal that a particular agent would be a poor and potentially unsafe social partner. We hypothesise that when encountering a social partner who has caused harm, we make act-based judgments about blame and punishment, and employ Partner Control to punish the perpetrators for causing harm to a victim. In contrast, for cases of moral impurity when there is no easily identifiable victim, we may instead rely more often on Partner Choice behaviours such as

avoidance, deselection and social group ostracization to avoid agents we deem to be morally bad or defective or who are perceived to have the potential to inflict future harm (without having caused harm right now). The reactive difference might exist because if harm has occurred one may imagine that Partner Control can "make right" the actual injustice. In other words, punishment may be felt to balance the scales using one harm as a response to another. In contrast, if there is a potential wrong in the future but no actual harm right now, there will be nothing necessarily to restore or reballence. Partner Choice, on the other hand, should vary based on the perceived character of the agent as a strategy to mitigate the potential for future harm. We aim to identify the categories of morality which are associated with one strategy over the other. We theorise that the function of being sensitive to the purportedly "harmless" foundations allows us to identify agents that would be potentially problematic so that we can avoid those agents. Across two preregistered studies we use a series of Harm and Purity vignettes to test the extent to which people react with either Partner Choice or Partner Control. We also ask participants to make judgments about these actions (Study 1: Moral Wrongness, Character Diagnosticity, discomfort around the violator even after they have been sufficiently punished (Discomfort); Study 2: Predictability of the violator (Predictability), understandability of motivations (Understandability), the predicted efficacy of punishment (Worry Before/After Punishment) the perceived harmfulness of the act (Harmful). We also explore the role of theoretically relevant demographic measures (Study 1: Political Orientation, Relational Robility and whether the participant lives in a rural or urban area (Urban/Rural). We test the theory that impurity will lead to Partner Choice reactions and harm will lead to Partner Control reactions, and we measure how demographic factors affect these overall reactions.

Experiment 1 - Partner Management & Moral Foundations

In Study 1 we test the extent to which Partner Control and Partner Choice reactions vary depending on the moral foundation being violated (including both first-person and preferred third-party reactions). We probed participants' attitudes about the violations (Moral Wrongness) and the violators (Character Diagnosticity, Discomfort). We test how people react to a selection of Purity and Harm vignettes while measuring theoretically relevant demographic measures, hypothesising that Purity violations will lead to Partner Choice and Harm violations will lead to Partner Control. We also hypothesise that individual differences in sensitivity to moral impurity are predicted by the tightness of the social group and how much opportunity it provides for Partner Choice. This study was pre-registered at: https://osf.io/vcu9g

Participants

We recruited 160 participants from the United States and Canada (aged between 20 and 83, M_{age} = 44.59, SD_{age} = 13.82, 51% male, 48% female) from Amazon Mechanical Turk in exchange for a small payment of $2.40. Participants were recruited based on the criteria that they must have had at least 50 previously accepted HITs and a prior approval rating above 90%. Cloud Research also enabled us to select a 50/50 split between self-identified left-wing vs right-wing individuals based on Cloud Research prescreening criteria (participants indicate whether they are liberal or conservative).

Procedure

Participants were split into two groups (ingroup condition / outgroup condition) where they read twelve vignettes (see table 2.1), taken from the Moral Foundations Vignettes (Clifford et al. 2015) and from the wider moral purity literature, six for Harm (example: "A student placing a thumbtack sticking up on the chair of another student") and six for Purity (example: "A man goes to the supermarket once a week and

buys a dead chicken. But before cooking the chicken, he has sexual intercourse with it"). All subjects saw all vignettes in random order. In the out-group condition people were told: "Please imagine that the people in these scenarios are complete strangers you have never met before." In the in-group condition, participants are told "Please imagine that the people in these scenarios are from your neighbourhood."

Table 2.1

Moral Violations

Violation	Harm / Purity	Name
You see a woman commenting out loud about how fat another woman looks in her jeans.	Harm	Fat
You see a zoo trainer jabbing a dolphin to get it to entertain his customers.	Harm	Dolphinjab
You see a student placing a thumbtack sticking up on the chair of another student.	Harm	Thumbtack
You become aware of someone conning an old man out of his savings	Harm	Conman
You learn that someone has broken into someone's house in order to steal their property	Harm	Theft
You find out someone has been lying about their co-worker to get them fired	Harm	Fired
A man goes to the supermarket once a week and buys a dead chicken. But before cooking the chicken, he has sexual intercourse with it.	Purity	Chickensex
You see a man searching through the trash to find women's discarded underwear.	Purity	Underwear
You see a man in a bar using his phone to watch people having sex with animals.	Purity	Animalsex
A man's dog was killed by a car in front of his house. He had heard that dog meat was delicious, so he cut up the dog's body and cooked it and ate it for dinner.	Purity	Dogfood
A brother and sister like to kiss each other on the mouth. When nobody is around, they find a secret hiding place and kiss each other on the mouth, passionately.	Purity	Siblingkiss
A man has been having sex with people in public places	Purity	Publicsex

For each vignette, participants marked on a 6-point Likert agreement scale what they felt this indicated about the moral character of the perpetrator (i.e., "This behaviour reveals a great deal about what this person is really like and their moral character"), the Moral Wrongness of the action (i.e., "This behaviour is morally wrong"), and their predicted Discomfort about being around the perpetrator after they would be punished (i.e., "I would feel uncomfortable around this person even after they had been appropriately punished"). Note that the moral character measure is a measure of diagnosticity (how much information is provided) rather than a measure of moral character itself (i.e the badness of the agent).

Partner Management reactions are measured with five slider questions, each with "Do Nothing" in the middle, the Partner Choice behaviour to the left and the Partner Control behaviour to the right. These questions were:"Would you generally AVOID this person, or APPROACH this person to express disapproval or DO NOTHING and not change behaviour towards this person either way"; "Would you make sure to NOT HAVE TO DEAL WITH this person, make sure they are PUNISHED for their actions or DO NOTHING and not change behaviour towards this person either way"; "Would you want your friends to generally AVOID this person, or want your friends to APPROACH this person in order to express disapproval or DO NOTHING and have your friends not change behaviour towards this person either way"; "Would you want this person to be generally NOT INCLUDED in the activities of your social group, have them be included in the social group but be MADE AN EXAMPLE OF, DO NOTHING and not change behaviour towards this person either way"; "Would you feel this person should generally be REJECTED, have to face RETRIBUTION, or face NO CONSEQUENCES". Note that these questions are framed as both first-person reactions (first two), preferences for third-person reactions (second two) and also abstract

preferences (final question). One reason for this is that participants may wish that the perpetrator is punished but, for safety, would rather that others would conduct the punishment.

Individual difference measures consist of participants' political orientation (using a 100 point slider from Liberal to Conservative), Relational Mobility using the standardised scale (Thomson et al. 2018) and the area in which they have lived with the question: “Would you describe the area for which you have lived most of you life or most identify with as largely urban and well-populated or rural with a small population” on a five-point scale from “very rural” to “very urban”.

Results

The results are split into four sections. The first section outlines Partner Management reactions (Partner Choice, Partner Control, Do Nothing) to the different Harm and Purity vignettes. Then the next section covers differences in judgments (Moral Wrongness, Character Diagnosticity, Discomfort around the violator) and how these judgments predict the Partner Management reactions. Finally, we test whether individual differences (Political Orientation, Relational Mobility, Urban vs Rural) mediate these judgments and reactions. Note: the ingroup-outgroup manipulation had no discernible effects on any dependent measure and therefore is left out of the analysis.

Both Bayesian and Frequentist analyses were preregistered. All frequentist mixed models initially specified all random slopes for all predictors and then random slopes were removed one by one to find the maximal converging model, which is recommended as best practice for linear mixed-effects models (Barr et al. 2013). Bayesian indices are able to quantify the evidence for and against effects, although they are more conservative than p-values. Frequentist analyses were performed using Jamovi v1.6 (Şahin and Aybek 2019), mixed models were calculated using the GAMLj module for Jamovi. Bayesian models (quantifying the levels of evidence for and against each predictor/model) were fitted using BayesFactor package with default settings (Morey et al. 2015). All the Bayesian mixed models had random slopes specified for all predictors. The bayesTestR package was used for calculating Bayes Factors of inclusion for comparing all possible models with the term against all equivalent models with that term removed rather than the term in the single model with the frequentist analyses (Makowski et al. 2019).

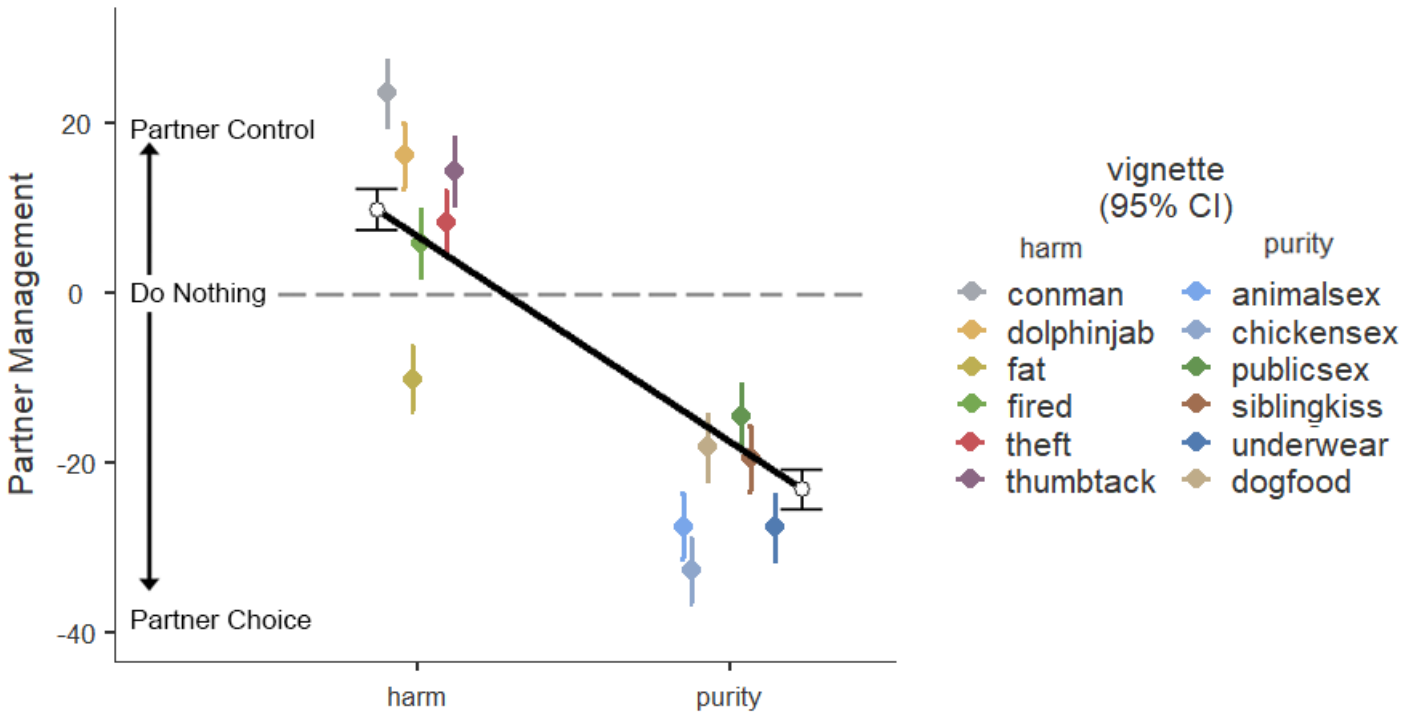
Partner Management Reactions of Harm and Purity Violations

Five questions were asked in order to measure Partner Management reactions between Partner Choice, Partner Control or doing nothing. A reliability analysis indicated this scale was highly reliable, Chronbach's $\alpha = 0.9$, and therefore the measure of Partner Management was the average of these answers.

In order to assess differences in Partner Management reactions between Harm and Purity violations, linear mixed models were calculated (Frequentist: random intercepts and random slopes for Foundation). Note that this is a simplification of the more complex preregistered models that are reported below, although results across models are consistent. The results indicate a significant main effect of Moral Foundation $F(1, 159) = 414$, $B = 33.02$, $SE = 1.62$, $p < 0.001$, $BF_{10} = 1.30 \times 10^{135}$. Purity violations consistently result in more Partner Choice than Harm Violations. The results for each vignette are illustrated in figure 2.1. Note that all Purty violations are on the Partner Choice side and all but one Harm violations are on the Partner Control side. The Harm vignette on the Partner Choice side is: *“You see a woman commenting out loud about how fat another woman looks in her jeans.”* Potential reasons for this are discussed in the Discussion section.

Figure 2.1.

Differences in Partner Management reactions for each vignette



To examine whether there is evidence that this pattern may generalise beyond the Harm and Purity violations used here, we performed an exploratory linear mixed model but included random intercepts for Vignette. The model indicated a significant effect of Foundation, suggesting that the results may generalise across other vignettes, $F(1, 11.2) = 34.1$, $B = 33.02$, $SE = 5.66$, $p < 0.001$, $BF_{10} = 1.30 \times 10^{135}$.

The Role of Moral Wrongness, Character Diagnosticity, and Discomfort Around the Perpetrator

It is possible that the difference in Partner Management reactions for Harm and Purity violators was driven by judgments of Moral Wrongness, Character Diagnosticity, or Discomfort. First we assessed whether there were differences in these judgments between Harm and Purity vignettes using linear mixed models (Frequentist: random intercepts and random slopes for Foundation). This was a simplification of the larger preregistered models that are calculated below, with results that were consistent.

Purity violations were judged to be significantly less Morally Wrong than Harm violations, although this difference was quite small, $F(1,159) = 44.2$, $B = 0.33$, $SE =0.05$, $p < 0.001$, $BF_{10} = 1.92 \times 10^{12}$. Both Harm and Purity violations were seen to be highly diagnostic of character, but the Purity violations were slightly less so, $F(1,159) = 34.7$, $B = 0.26$, $SE = 0.04$, $p < 0.001$, $BF_{10} = 1.49 \times 10^{7}$. Purity violations also resulted in more Discomfort although Baysian evidence is anecdotally against this effect, $F(1,159) = 5.32$, $B = 0.11$, $SE = 0.05$, $p = 0.022$, $BF_{10} = 0.909$. See figure 2.2 for plots of each judgement. Note that across vignettes, judgments are spread relatively evenly.

In order to assess the extent to which these differences explain the different Partner Management reactions for Harm and Purity violations, pre registered linear mixed models were calculated (Frequentist: random intercepts and random slopes for Foundation), see table 2.2 for results.

Harm violations lead to Partner Control, the strength of which is associated with how bad the act was, whereas Purity violations lead to Partner Choice which is associated with how bad the agent is (Character Diagnosticity and Discomfort), see figure 2.2.

Note that Character Diagnosticity and Moral Wrongness and Discomfort are highly correlated with each other, Moral Wrongess with Character, $r(1918) =0.61$, $p < 0.001$; Moral Wrongess with Discomfort, $r(1918) = 0.56$, $p < 0.001$; Discomfort with Character $r(1918) = 0.63$, $p < 0.001$. This means it is difficult to isolate how they may drive each other and in turn drive Partner Management judgments.

Table 2.2

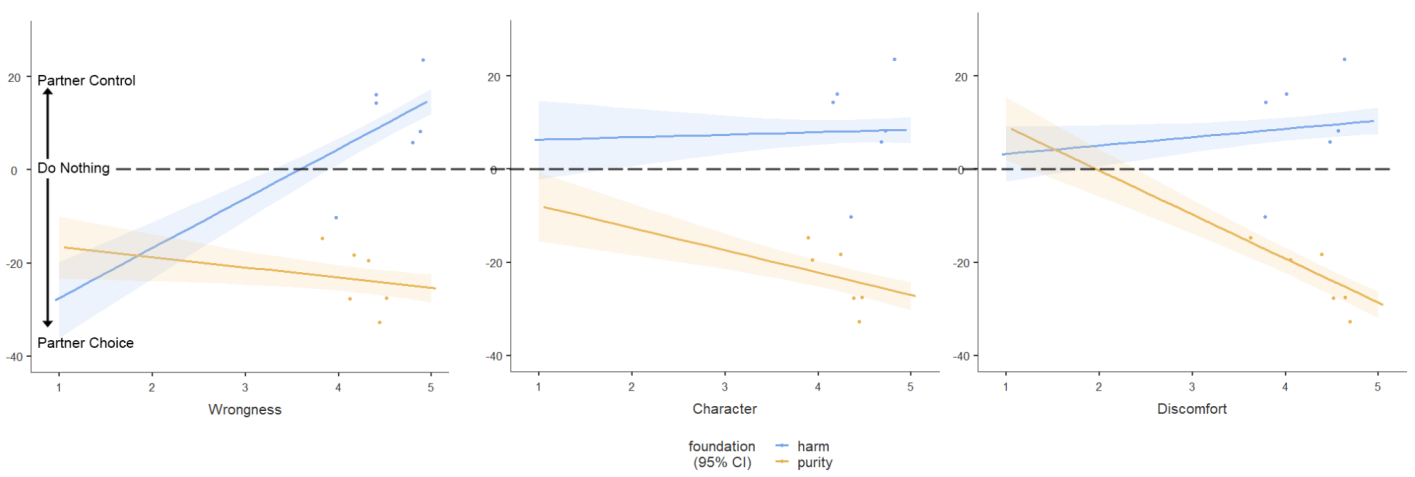
Statistics and evidential indices for each judgement predicting Partner Management.

	Estimate	F	df	p	Inclusion BF
Foundation	-10.08	218.08	359.31	< .001***	2.61×10^{118} △△△
Wrongness	-25.87	50.56	1850.84	< .001***	1.77×10^{04} △△△
Character	7.47	0.43	1825.07	0.511	0.128▽
Discomfort	0.71	30.28	1877.62	< .001***	1.01×10^{6} △△△
Foundation ✻ Wrongness	-4.87	33.78	1798.63	< .001***	466.37 △△△
Foundation ✻ Character	-12.00	6.50	1513.82	0.011**	1.37◦
Wrongness ✻ Character	-5.24	10.17	1838.73	0.001***	6.05△
Foundation ✻ Discomfort	2.52	8.33	1739.11	0.004**	13.77△△
Wrongness ✻ Discomfort	-4.97	8.36	1784.70	0.004**	0.294▽
Character ✻ Discomfort	2.34	1.25	1827.44	0.264	0.269▽
Foundation ✻ Wrongness ✻ Character	0.81	3.79	1823.02	0.052*	0.420◦
Foundation ✻ Wrongness ✻ Discomfort	-3.06	5.78	1787.62	0.016**	0.516◦
Foundation ✻ Character ✻ Discomfort	-3.87	3.99	1880.03	0.046*	132.98 △△△
Wrongness ✻ Character ✻ Discomfort	-2.92	9.32	1797.02	0.002**	0.184▽
Foundation ✻ Wrongness ✻ Character ✻ Discomfort	1.39	3.48	1801.75	0.062	1.42◦

*Note: * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, △ indicates substantial evidence $BF_{10} > 3$, △△ indicates strong evidence $BF_{10} > 10$, △△△ indicates extreme evidence $BF_{10} > 100$, ▽ indicates substantial evidence for the null $BF_{10} < 0.33$, ◦ indicates test was insensitive BF_{10} between 0.33 & 3.*

Figure 2.2.

Moral Judgments Predicting Partner Management Reactions



Note: Points represent empirical averages per vignette

In order to assess whether reactions to Purity and Harm violations are fully explained by the probed judgments (Wrongness, Character Diagnosticity, and Discomfort), three models were compared. A Foundations Model containing only Moral Foundation (whether the violations was Harm or Purity), a Judgments Model containing Wrongness, Character Diagnosticity, and Discomfort and a Full Model containing both these sets of predictors. The Full Model containing both sets decisively outperformed the Foundations model $BF_{10} = 4.18 \times 10^{31}$; and outperformed the Judgments model $BF_{10} = 8.20 \times 10^{145}$. In addition, the Foundation-only model outperformed the Judgments model $BF_{10} = 1.96 \times 10^{114}$. Taken together these results indicate that the different reactions elicited from Harm and Purity violations can only be partly explained by the moral judgments assessed in this study.

To determine whether the Moral Foundation is decisive in predicting whether people engage in Partner Choice or Partner Control, we plotted model predictions from the data when the terms for “Purity” and

"Harm" are switched and everything else is kept the same. Results confirmed that switching the Harm and Purity terms predominantly switches the predicted reactions from Partner Choice to Partner Control and vice versa.

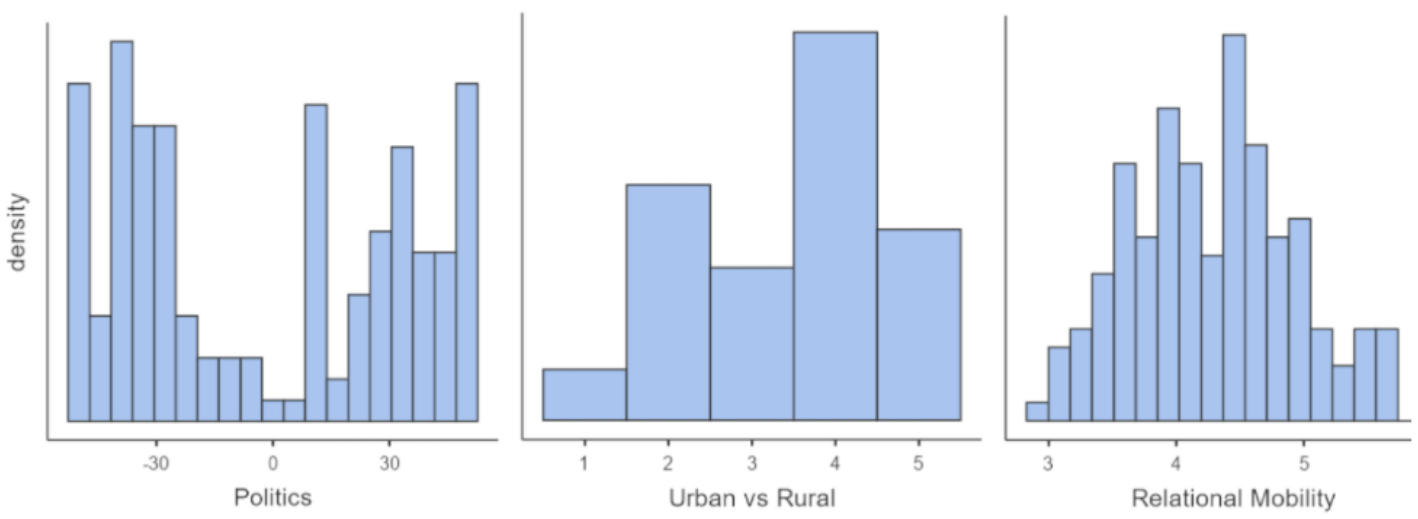
Despite Moral Wrongness, Character Diagnosticity, and Discomfort all affecting the Partner Management strategy to some degree, they are not able to fully explain why Purity violations result in Partner Choice and Harm violations in Partner Control.

Individual Differences

We used pre-screening procedures to ensure an even split between left-wing and right-wing participants. They reported their political orientation on a scale from -50 (most Liberal) to 50 (most Conservative), M = - 1.93, SD = 35.75. Participants were also asked whether they lived in a more urban or rural area where 1 = Very Rural and 5 = Very Urban, M = 3.42, SD = 1.18. Finally participants complete the relational mobility scale M = 4.32, SD = 0.66, see figure 2.3.

Figure 2.3

Distribution of Participants' Political Orientation, Relational Mobility & Urban vs Rural Living Situation



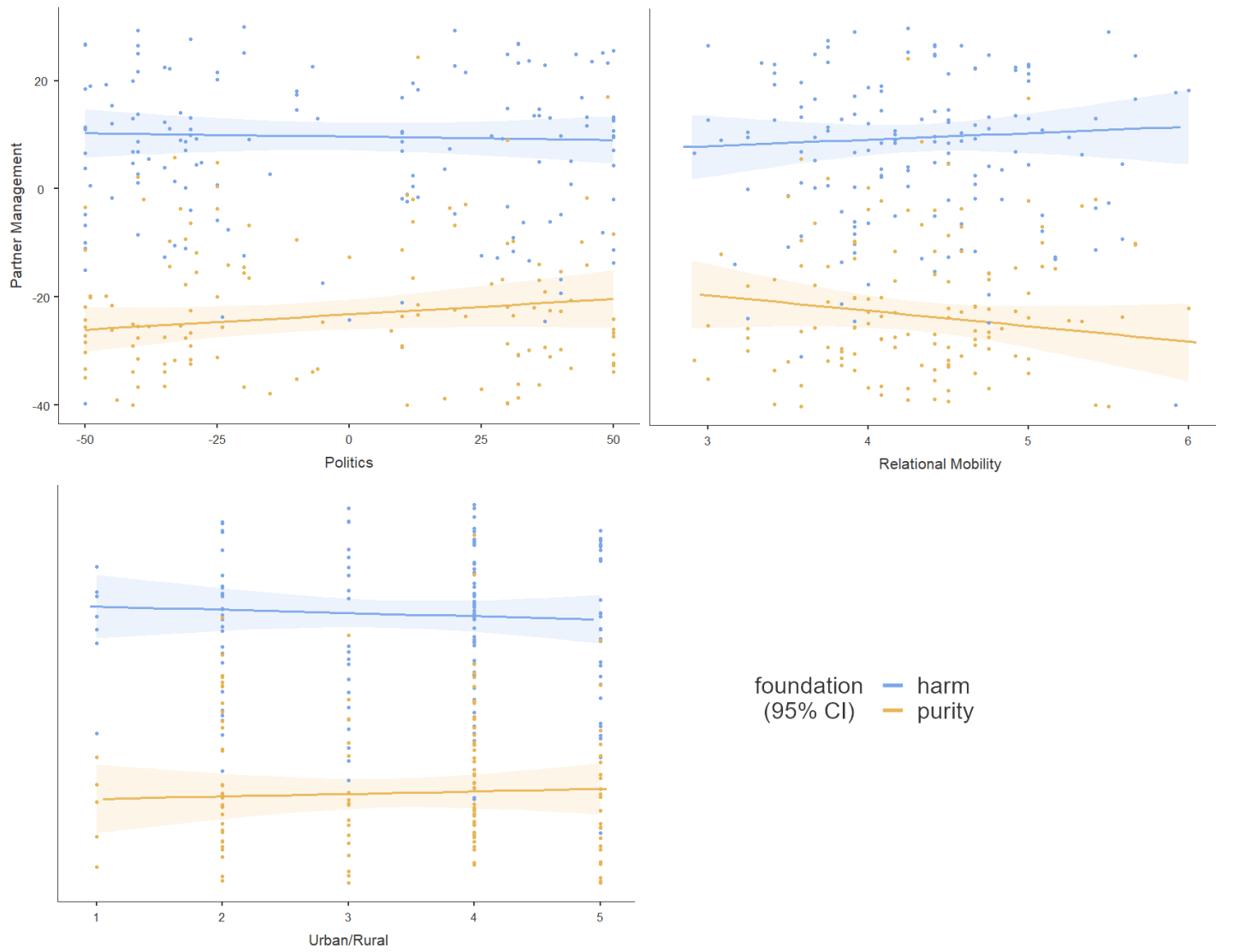
We expected that these three demographic measures, (Political Orientation, Relational Mobility, and Urban/Rural) may be correlated with each other but surprisingly, results indicate that this was not the case. Political Orientation - Relational Mobility, r(158) = 0.03, p = 0.716, , BF_{10} = 0.11; Political Orientation - Rural/Urban p = 0.061, r(158) =-0.15, BF_{10} = 0.57, Relational Mobility - Rural/Urban, r(158) =-0.07, p = 0.348, BF_{10} = 0.15. Because of this, for the following analyses, the demographic measures are assessed in the same model rather than in separate models.

Pre-registered mixed models (Frequentist: random intercepts and random slopes for Foundation Politics and Rural-Urban) are the same as the models above for predicting Partner Management reactions but with the demographic measures included as predictors. Moral Foundation remained significant, F(1,156.07) = 418.67, B = -33.02, SE = 1.04, p < 0.001, $BF_{10} = 1.39 \times 10^{135}$. However, all demographic measures and their interaction with Moral Foundation were non-significant; Politics, F(1,94.46) = 0.42, B = 0.02, SE = 0.03, p = 0.71, BF_{10} = 0.104; Relational Mobility, F(1,69.18) = 0.31, B = -0.85, SE = 1.10, p = 0.580, BF_{10} = 0.219; Rural/Urban, F(1,37.83) < 0.01, B = -0.06, SE = 1.54, p = 0.949, BF_{10} = 0.076; Foundation by Politics, F(1,156.72) = 2.24, B = 2.43, SE = 1.62, p = 0.137, BF_{10} = 0.403; Foundation by Relational Mobility, F(1,153.88) = 2.85, B = -4.08, SE = 2.42, p = 0.094. BF_{10} =1.05; Foundation by Rural/Urban, F(1,152.33) = 0.59, B = 1.06, SE = 1.37, p = 0.442, BF_{10} = 0.127

This indicates that even across the spectrum of Political Orientation, Relational Mobility and Urban vs Rural living environment, Purity violations reliably result in more Partner Choice than Harm Violations which reliably result in more Partner Control, see figure 2.4.

Figure 2.4.

Individual Differences Predicting Partner Management Reactions



Mixed models were calculated to assess whether individual differences predicted differences in judgments between Harm and Purity violations. These preregistered models are the same as the models above for judgments of Harm and Purity violations but with the individual differences measures included as predictors.

For models predicting Moral Wrongness (Frequentist: Random intercepts and slopes for foundation), there were was significant effects for Relational Mobility, $F(1,156) = 5.06$, $B = 0.12$, $SE = 0.05$, $p = 0.026$, $BF_{10} = 278.61$, and there was a Foundation by Politics interaction, $F(1,156) = 15.44$, $B = 19$, $SE = 0.05$, $p < 0.001$, $BF_{10} = 1870$. However, there was no significant main effects of Foundation, $F(1,156) = 0.3$, $B = -19$, $SE = 0.36$, $p = 0.586$, $BF_{10} = 2.44 \times 10^{12}$ (note that Bayesian evidence is strongly in favour of this effect when averaging across all possible models); Political Orientation $F(1,156) = 0.23$, $B = 0.02$, $SE = 0.04$, $p = 0.634$, $BF_{10} = 0.131$; Urban/Rural, $F(1,156) = 0.12$, $B = -0.01$, $SE = 0.03$, $p = 0.734$, $BF_{10} = 0.134$; Urban/Rural by Foundation interaction $F(1,156) = 0.64$, $B = 0.03$, $SE = 0.04$, $p = 0.423$, $BF_{10} = 0.207$ and Relational Mobility by Foundation interaction, $F(1,156) = 0.59$, $B = -0.06$, $SE = 0.07$, $p = 0.445$, $BF_{10} = 0.188$

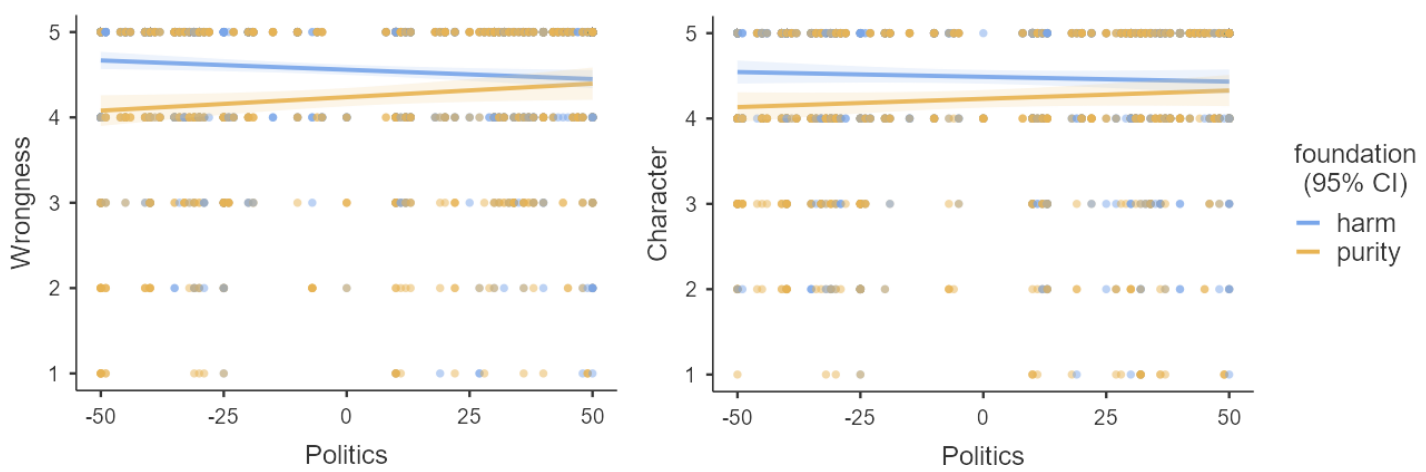
The interaction between Politics was consistent with the classic findings from MFT (Kivikangas et al. 2020; Graham et al. 2009) that conservatives care more about purity and less about harm than liberals. In addition, contrary to expectations higher Relational Mobility led to stronger judgments of Moral Wrongness. This is discussed further in the Discussion section.

The same analyses were run predicting Character Diagnosticity (Frequentist: Random intercepts and slopes for Foundation and Urban-Rural). Results indicated a significant main effects of Foundation, $F(1, 156.01) = 35.45$, $B = -0.26$, $SE = 0.04$, $p < 0.001$, $BF_{10} = 2.22 \times 10^{7}$, and Relational Mobility, $F(1, 152) = 11.51$, $B = 0.2$, $SE = 0.06$, $p < 0.001$, $BF_{10} = 2.42 \times 10^{6}$, and also a Foundation by Politics interaction, $F(1, 152) = 6.27$, $B = 0.11$, $SE = 0.04$, $p = 0.013$, $BF_{10} = 3.66$. However, there was no significant effect of Politics, $F(1, 139.16) = 0.34$, $B = 0.02$, $SE = 0.04$, $p = 0.561$, $BF_{10} = 0.126$, Urban-Rural, $F(1, 87.35) = 0.71$, $B = -0.03$, $SE = 0.04$, $p = 0.400$, $BF_{10} = 2.00$, and there was no Foundation by Relational Mobility interaction, $F(1, 155.72) = 0.29$, $B = -0.04$, $SE = 0.07$, $p = 0.593$, $BF_{10} = 0.109$ or a Foundation by Urban Rural interaction, $F(1, 156.03) = 0.01$, $B = 0.01$, $SE = 0.04$, $p = 0.941$, $BF_{10} = 0.115$.

Higher Relational mobility resulted in stronger Character Diagnosticity judgments and the more Conservative the stronger Character Diagnosticity for Purity and the weaker Character Diagnosticity for Harm, see figure 2.5.

Figure 2.5.

Interaction between Moral Foundation and Political Orientation for Judgments of Wrongness & Character Diagnosticity



The same analyses were run predicting how uncomfortable the participant predicted they would feel, even after the violator was adequately punished (Frequentist: Random intercepts and slopes for Foundation). Results indicated that Purity violations elicited more Discomfort than Harm violations, although Bayesian evidence was anecdotally against this effect, $F(1, 156) = 5.40$, $B = 0.1$, $SE = 0.05$, $p = 0.021$, $BF_{10} = 0.896$.

However, there were no significant main effects of, Politics, $F(1, 156) = 0.07$, $B = -0.1$, $SE = 0.04$, $p = 0.790$, $BF_{10} = 0.092$; Relational Mobility, $F(1, 156) = 2.87$, $B = 0.11$, $SE = 0.06$, $p = 0.092$, $BF_{10} = 4.84$; Urban-Rural, $F(1, 156) = 0.04$, $B = 0.1$, $SE = 0.04$, $p = 0.838$, $BF_{10} = 0.086$; nor Foundation by Politics Interaction, $F(1, 156) = 3.06$, $B = 0.08$, $SE = 0.05$, $p = 0.082$, $BF_{10} = 0.681$; Foundation by Relational Mobility Interaction; $F(1, 156) = 0.62$, $B = -0.6$, $SE = 0.07$, $p = 0.431$, $BF_{10} = 0.109$ and Foundation by Urban-Rural Interaction, $F(1, 156) = 1.14$, $B = -0.04$, $SE = 0.04$, $p = 0.287$, $BF_{10} = 0.241$.

Note, despite the non-significance of Relational Mobility in the large model, Bayesian evidence was in favour of this effect, where stronger Discomfort was elicited from those with higher Relational Mobility. Therefore, a simple mixed model was calculated (Frequentist: random intercepts) but it found no significant main effect of Relational Mobility $F(1, 158) = 2.84$, $B = 0.10$, $SE = 0.06$, $p = 0.094$, $BF_{10} = 4.84$; Urban-Rural, $F(1, 156) = 0.04$, $B = 0.1$, $SE = 0.04$, $p = 0.838$, $BF_{10} = 4.54$, although Bayes factors still indicated some weak evidence in favour of this compared to the null.

Discussion

As predicted, harm leads to Partner Control and impurity to Partner Choice. When reacting to Harm violations with Partner Control, the reaction is more sensitive to how morally wrong the action was judged to be. In contrast, when reacting to Purity violations with Partner Choice, the reaction is more sensitive to the perceived discomfort around the violator, even after they had been punished. To some extent the same may be said for character diagnosticity but the evidence is less clear here as Bayesian evidence was only anecdotal for the two-way interaction. Although Harm violations, in general, were felt to be more morally wrong and more diagnostic of moral character while Purity violators produced more discomfort, these patterns were spread fairly evenly between each vignette.

Note that one Harm violation was predominantly met with Partner Choice. This vignette was: "You see a woman commenting out loud about how fat another woman looks in her jeans." One possibility is that this harm violation was not perceived to be particularly harmful compared to the others but that it revealed enough about the perpetrator's character that indicated they ought to be avoided or ostracised. Another possibility is that some participants may have interpreted that this happened out of earshot of the woman and therefore it is not clear that there is a harmed victim in this instance. This is explored further in Experiment 2.

Surprisingly, the individual differences measures did not predict differences in Partner Management but they did predict differences in judgement. Moral Wrongness judgments matched similar findings from Moral Foundations theorists, (Kivikangas et al. 2020; Graham et al. 2009) where conservatives care more about Purity and less about Harm than liberals. This pattern was replicated with judgments of character diagnosticity where conservatives compared to liberals felt Purity violations were more revealing about character than Harm.

Surprisingly the violations, in general, were met with stronger judgments for moral wrongness, character diagnosticity and (although the evidence was weaker) discomfort, when participants had higher Relational Mobility. One possible explanation for this is that when one is able to choose between potential social partners, it is beneficial to lean on stronger moral judgments in order to decide between agents in the biological market. However, this is in conflict with the idea that lower relational mobility societies have stronger cultural norms, for example they are more likely to be honour cultures (Thomson et al. 2018). The present finding is hard to explain and goes against our predictions. This should be explored further in future work that has cross-cultural samples.

Higher moral wrongness leads to more Partner Control and conservatives have stronger moral wrongness judgments for Purity violations. However, remarkably this seems to have had no significant bearing on what Partner Management strategy conservatives prefer. This may be because the differences between conservatives and liberals on moral wrongness judgments is simply not large enough to have a significant effect on Partner Management preferences. However, conservatives also felt that Purity violations were more diagnostic of character which predicts greater preferences for Partner Choice. Since the differences in sensitivity to moral impurity between liberals and conservatives can be difficult to isolate (Kivikangas et al. 2020), this question should be explored more thoroughly in future research.

The moral judgments participants made (Wrongness, Character Diagnosticity & Discomfort) can help partly explain differences in Partner Management but not entirely as including Moral Foundation in the model still drastically increased predictive performance and switching the terms for Moral Foundations (while keeping the other judgments constant) was found to predominantly switch the model predictions from Partner Control to Partner Choice and vice-versa. This indicates that the moral wrongness, character

diagnosticity and discomfort moderate the strength of the selected Partner Management strategy but they are not decisive in which strategy is preferable. Other aspects of why Purity violations differ from Harm violations must therefore be responsible for this difference.

The purpose of Experiment 2 is first to confirm that Harm is indeed more associated with Partner Control by using new dependent measures and second to assess the effect of a further set of judgments and emotional reactions. In addition, Experiment 2 will dissociate between 1st party and 3rd party reactions to test whether participants would be biassed away from 1st party punishment and whether this bias is specific to particular kinds of violator.

Experiment 2 - Partner Management Mediating Factors

This second study tests the extent to which the difference in Partner Management reactions can be explained by a wider set of judgments and emotional reactions. People react to the same selection of Purity and Harm violations, though there is one swapped vignette- conman is replaced with a murder "You learn that someone intentionally killed his acquaintance because he was jealous of them"- to ensure that this experiment contains a particularly bad harm violation. We measure the extent to which each violation signals the predictability of the violator, their perceived motivations, worry about the violator causing future harm both before and after punishment and the perceived harmfulness of the act. In addition, we probe the emotions elicited by the violation, the violator and self-directed emotions if the participant imagines themself as the violator.

We also add an alternative measure of Partner Management using two three-way forced-choice questions. This allows us to test both the chance of reacting and also, given that there was a reaction, was it Partner Choice or Partner Control. Across the two three-way forced-choice measures, one probes the participants' Partner Management reactions and the other probes their preferences over 3rd-party reactions. The study pre-registration can be found here: https://osf.io/4h5rf

Participants

We recruited 117 participants (aged between 18 and 21, M_{age} = 18.66, SD_{age} = 0.56, 14% male, 84% female) from Warwick University's psychology undergraduate student subject pool. Participants received course credit taking part in the experiment.

Procedure

The design follows the same procedure as Study 1 excluding individual differences and the swapped vignette discussed above. In addition, we measure Partner Management slightly differently. We use two three-way forced choice questions, 1st Person Reaction: "I would rather: Avoid Them/Punish Them/Do nothing" and 3rd Party Reaction Preference: "I would rather people I know: Avoid them/Have them Punished/Do nothing." To check the consistency of this measure with the slider questions in the previous study, we also include one slider question, "Would you feel this person should generally be REJECTED, have to face PUNISHMENT or face NO CONSEQUENCES." Participants are once again asked to judge the moral wrongness of each violation but they are also asked, "How would you judge this person's predictability?"; "How difficult is it to understand the motivations for this person's actions (e.g., understand why he acted the way he did)?"; "How much harm would you say this action caused?" In addition, to assess predicted punishment efficacy they are asked "How worried would you be about this person causing harm in the future?" and "[…] causing harm in the future even if this person was sufficiently punished?"

To measure emotional reactions to each violation, participants are asked two questions, “Which emotion best represents how you feel about this [action/person]” where they respond with a slider on a scale between disgust and anger. They are also asked, “If you had done this, which emotion best represents how you would feel?” with a slider on a scale between shame and guilt. Finally, the new harm vignette that was used, in order to include a harm that was particularly severe, “You learn that someone intentionally killed his acquaintance because he was jealous of them.”

Results

As with Experiment 1, frequentist mixed-model regressions were calculated to assess how different foundations lead to different Partner Management reactions, however dependent measures were on categorical forced-choice scales. Therefore two logistic mixed model regressions were pre-registered, first for predicting the chance of Reacting vs Doing Nothing and second for predicting what reaction was chosen between Partner Choice and Partner Control (i.e. when *Do Nothing* is not chosen). Frequentist analyses are run the same way with the same software as Experiment 1 but Bayesian logistic mixed-models are calculated with BRMS (Bürkner 2017) since the BayesFactor package used for Experiment 1 does not allow for this kind of analysis.

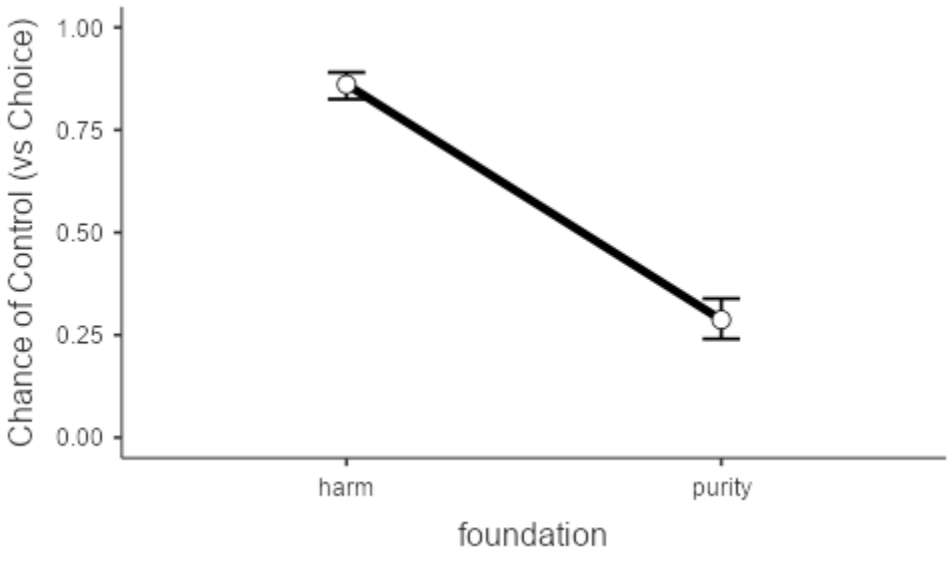
Partner Management Reactions of Harm and Purity Violations

A logistic mixed model regression (frequentist: random intercepts and random slopes for Foundation) predicting whether participants React vs Do Nothing revealed no significant difference between Harm and Purity violations with Bayesian analysis indicating evidence against this $\chi2(1) = 2.64$, $B = 0.49$, $SE = 0.3$, $p = 0.104$, $BF_{10} = 0.22$. In fact, participants rarely chose the “Do Nothing” option (Harm: 0.08%, Purity: 0.1%) indicating that participants had strong preferences to react to each moral violation.

For all reactions, a logistic mixed model regression predicting Reaction Type (Partner Choice vs Partner Control) (frequentist: random intercepts and random slopes for Foundation) revealed a significant effect of Moral Foundation between Harm and Purity Violations. $\chi2(1) = 283.49$, $B = 2.73$, $SE = 0.16$, $p < 0.001$, $BF_{10} = 1.15 \times 10^{36}$. As predicted, Harm violations were significantly more likely to result in Partner Control and Purity violations were significantly more likely to result in Partner Choice , see figure 2.6.

Figure 2.6

Chance of Partner Control between harm and purity violations



As with Experiment 1 we also included a single continuous measure of Partner Management. A Mixed model regression predicting Partner Management reactions with random intercepts and random slopes

for Foundation demonstrated a significant difference between Harm and Purity violations, $F(1,116) = 318.04$, $B = 33.17$, $SE = 1.86$, $p < 0.001$, $BF_{10} > 1000$, see figures 3.7 & 3.8. T-Tests assessed whether Partner Management reactions for Purity violations was significantly below zero (towards Partner Choice) and results indicated this was indeed the case, $t(701) = 2.75$, $p = 0.003$, $BF_{10} = 5.63$. Harm was significantly above zero (towards Partner Control), $t(701) = 34.3$, $p < 0.001$, $BF10 = 1.91 \times 10^{148}$. Therefore, the continuous measure corroborates both Experiment 1 and the strong results of the forced-choice analysis, though judgments appeared overall more in favour of Partner Control, perhaps due to the sample for Experiment 2 being young psychology undergraduates instead of the general population.

Note also that unlike Experiment 1 the *fat* violation is met with Partner Control. This may be due to the forced-choice nature of the dependent variable.

Figure 2.7

Linear measure of Partner Management between harm and purity violations

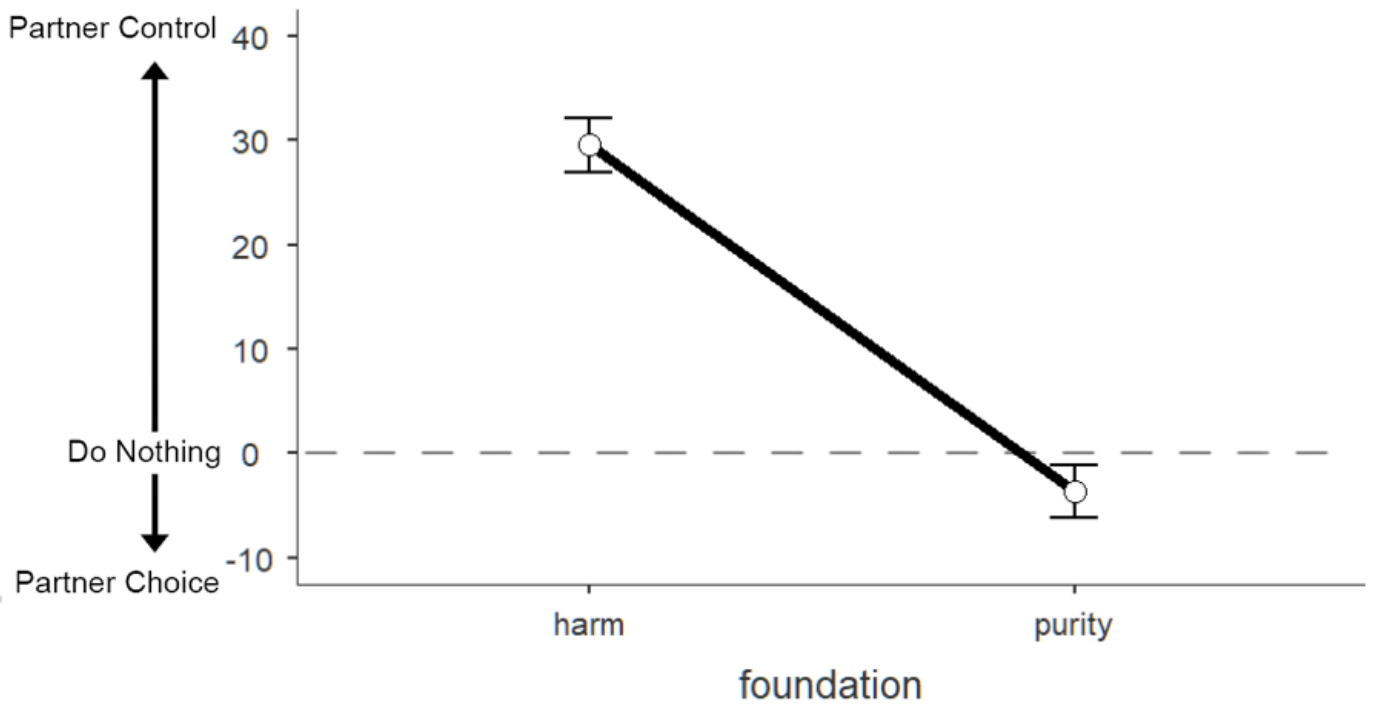
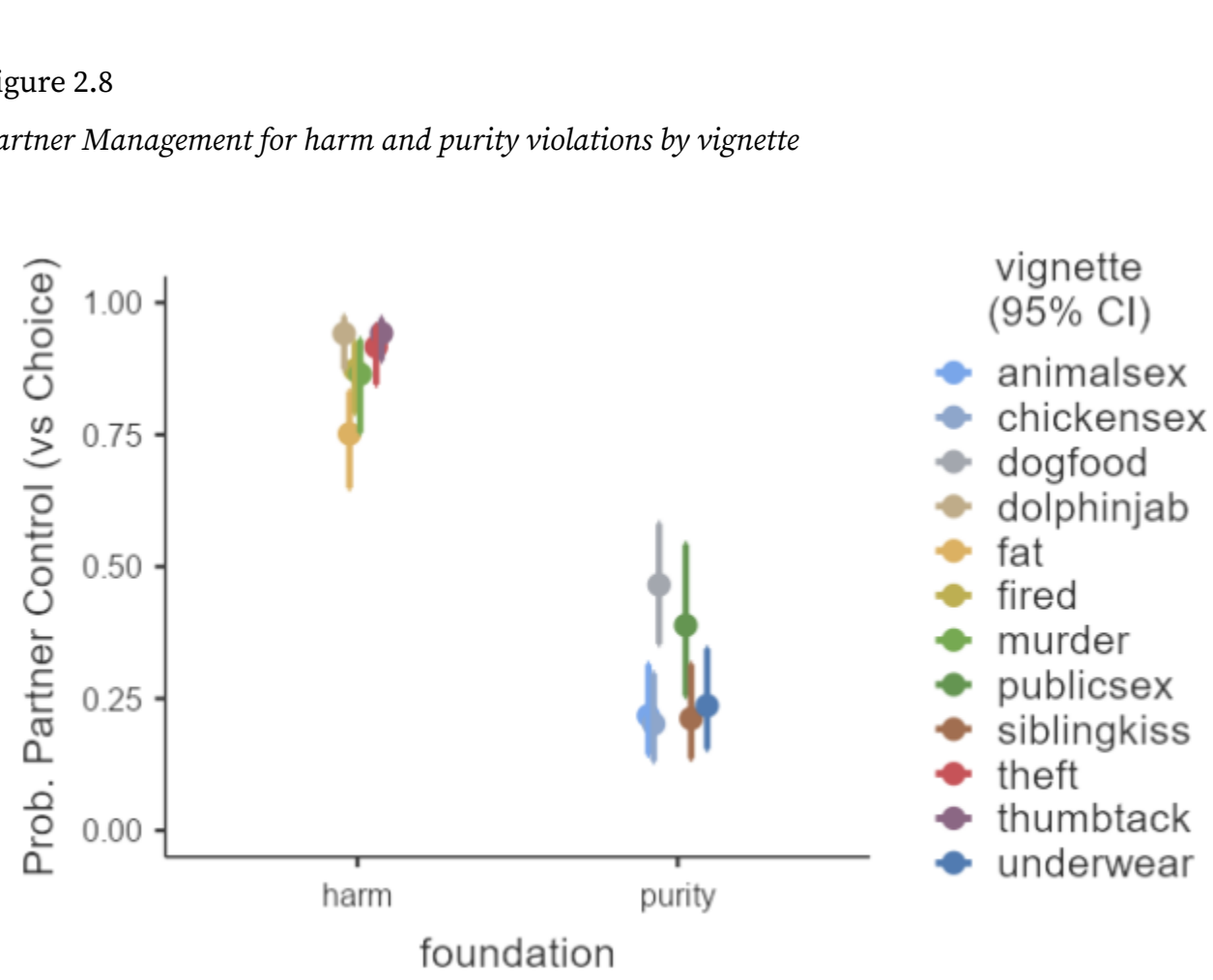


Figure 2.8

Partner Management for harm and purity violations by vignette



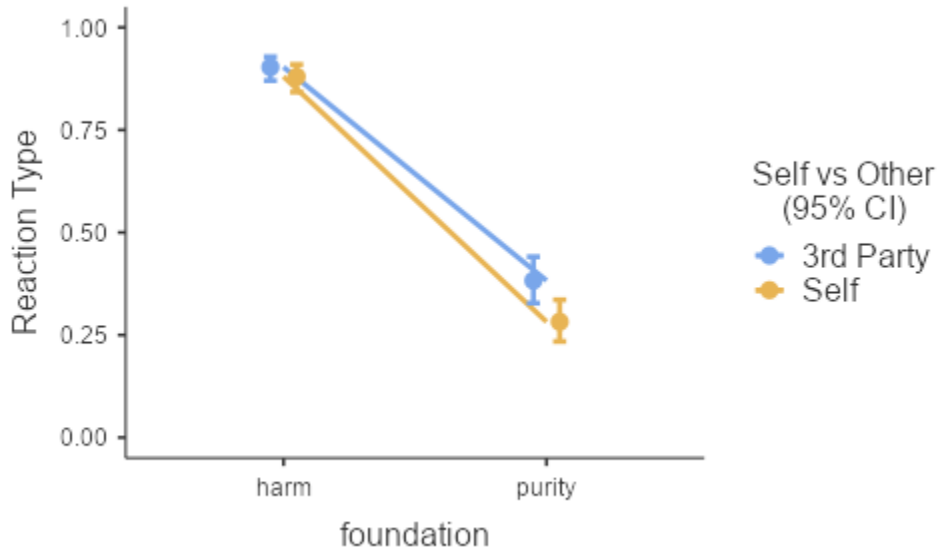
Self versus Third Party Reactions

To assess whether participants were worried about the dangers of punishing violators themselves we also assessed what participants would rather 3rd-parties do. As with before, for predicting whether or not "Do Nothing" is chosen, a logistic mixed model (frequentist: random intercepts and random slopes for Foundation) indicated a main effect of Foundation and Self vs 3rd Party, Foundation $\chi2(1) = 6.92$, $B = 0.78$, $SE = 0.3$, $p = 0.009$, $BF_{10} = 1.72$; Self vs 3rd Party $\chi2(1) = 5.02$, $B = 0.33$, $SE = 0.15$ $p = 0.025$ $BF_{10} = 0.289$. However, strong conclusions should not be drawn from this as Bayesian analyses indicate evidence against an effect of Self vs Other and that the test was insensitive for Foundation. In addition, there was no significant interaction effect $\chi2(1) = 1.07$, $B = 0.31$, $SE = 0.3$, $p = 0.301$, $BF_{10} = 1.72$ with Bayesian analysis indicating only anecdotal evidence for this effect.

For what reaction participants prefered, results indicated a main effect of Foundation and Self vs 3rd Party, Foundation $\chi2(1) = 297.37$, $B = 2.82$, $SE = 0.16$, $p = 0.001$, $BF_{10} = 2.55 \times 10^{34}$; Self vs 3rd Party $\chi2(1) = 10.73$, $B = 0.35$, $SE = 0.11$, $p < 0.01$, bf $= 13.99$. Purity violations resulted in significantly more Partner Choice, as expected, and Partner Control was slightly more favoured for third parties, see figure 2.9. There was no significant interaction effect, $\chi2(1) = 1.02$, $B = 0.21$, $SE = 0.21$, $p = 0.313$, $BF_{10} = 0.101$ with Bayesian analysis indicating evidence against this. This indicates that people were not more worried about enacting punishment for specific violators (i.e. Harm vs Purity violators) compared to having a third party do it.

Figure 2.9

1st person vs 3rd party reactions



The Roles of Moral Wrongness, Predictability, Understandability, Harmfulness, Future Harm, and Emotional Reaction

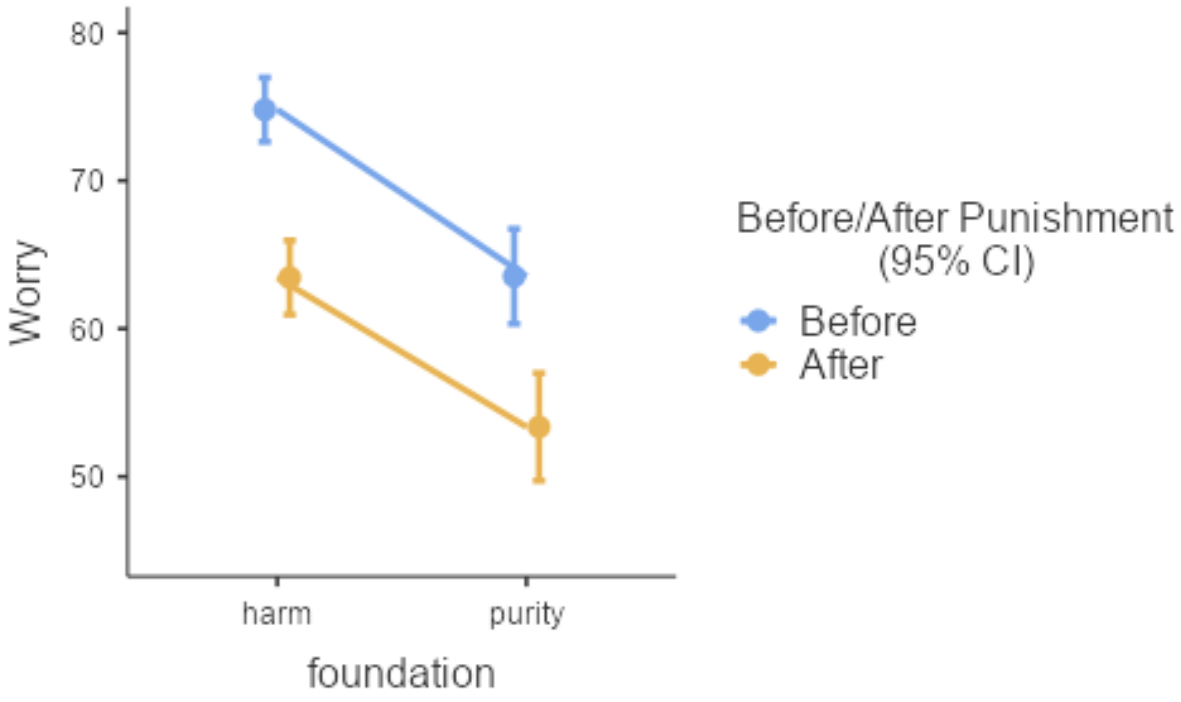
Four linear mixed models (predicting: Moral Wrongness, Predictability, Understandability and Harmfulness) were calculated. Moral Wrongness, $F(1,1286) = 42.03$, $B = 7.24$, $SE = 1.12$, $p < 0.001$, $BF_{10} > 1000$; Predictability, $F(1,1286) = 95.39$, $B = 11.94$, $SE = 1.22$, $p < 0.001$, $BF_{10} > 1000$; Understandability, $F(1,1286) = 149.16$, $B = 16.22$, $SE = 1.33$, $p < 0.001$, $BF_{10} > 1000$; Harmfulness, $F(1,1286) = 716.5$, $B = 31.4$, $SE = 1.17$, $p < 0.001$, $BF_{10} > 1000$. Harm violations were rated as significantly worse and more harmful than Purity violations, Purity violators were rated as significantly less predictable than Harm violators and had significantly less understandable motivations

Participants were asked both how worried they would be about future harm from the perpetrator, and also how worried they would be even after the perpetrator was sufficiently punished. Mixed models (frequentist: random intercepts and random slopes for Foundation and Before/After) indicated that

participants were significantly more worried about Harm violators causing future harm $F(1,467.88) = 156.54$, $B = 10.76$, $SE = 0.86$, $p < 0.001$, $bf > 1000$, and that people would be significantly less worried after punishing the perpetrator $F(1,116) = 50.37$, $B = 10.68$, $SE = 1.50$, $p < 0.001$, $bf > 1000$ see figure 2.10. However, there was no significant interaction between Worry Type and Foundation, with Bayesian analysis indicating evidence against this effect, $F(1,2572) = 0.51$, $B = 1.17$, $SE = 0.86$, $p = 0.474$, $BF_{10} = 0.069$. Contrary to what was predicted, this indicates that participants have similar feelings regarding the efficacy of Partner Control for Purity violators compared to Harm violators, see figure 2.10.

Figure 2.10

Worry about being around the violator before and after punishment



A mixed-model predicting Emotional Reactions (frequentist: random intercepts and random slopes for Foundation) indicated a main effect of Foundation, Emotion Type (disgusts vs anger at the perpetrator and at the action, shame vs guilt at self) and their interaction, Foundation $F(1,115.8) = 963.54$, $B = 49.46$, $SE = 1.59$, $p < 0.001$, $bf > 1000$; Emotion Type $F(2,3964.41) = 4.70$, $B = 2.72$, $SE = 1.06$, $p = 0.09$, $BF_{10} = 0.182$, Interaction $F(2,3964.41) = 5.32$, $B = 5.68$, $SE = 1.52$, $p = 0.005$, $BF_{10} = 0.588$. As expected, harm violations resulted in more anger/guilt and purity violations resulted in more disgust/shame, see figure 2.11. However, strong conclusions should not be drawn from the main effect of Emotion Type nor the interaction with Foundation as Bayesian analyses suggest there is evidence against this (emotion type) or that the test was insensitive (interaction).

Figure 2.11

Differences in Emotional Reaction to Harm and Purity Vignettes

Note. > 0 is more Anger/Guilt) and < 0 is more Disgust/Shame.

A reliability analysis of the three kinds of Emotional Reaction showed they are highly reliable, Cronbach's α = 0.8 and therefore they were averaged for subsequent analyses assessing Partner Management.

In order to assess the role of perceived Moral Wrongness, Predictability (of the perpetrator), Understandability (of their motivation), Harmfulness (of the act), Worry (about future harm), and Emotion (elicited from the act/person) on predicting Partner Management reactions, a logistic mixed model was calculated (frequentist: random intercepts and random slopes for Worry and Understandability). The results are presented in table 2.3. The more morally wrong and harmful the act is perceived to be, and more worried about future harm and angry the participant is, the greater the likelihood of Partner Control, whereas the less understandable and predictable the violator and more disgusted the participant feels the greater chance of Partner Choice, see figure 2.12.

Table 2.3.

Results of all judgments predicting Partner Management

Fixed Effect Omnibus tests

	χ^2	β	df	p
Wrongness	36.11	0.04	1	<0.001***
Predictability	3.71	-0.01	1	0.056
Understandability	10.59	-0.01	1	<0.001***
Harmfulness	5.31	0.01	1	0.078
Foundation	21.6	-0.39	1	<0.001***
Emotion	31.97	0.03	1	<0.001***
Worry	1.56	0.01	1	0.014*
Foundation ✻ Wrongness	5.15	0.03	1	0.015*
Foundation ✻ Worry	0.00	3.67e-4	1	0.469
Foundation ✻ Harmfulness	0.08	-0.00	1	0.804
Foundation ✻ Understandability	0.00	-4.74e-4	1	0.976
Foundation ✻ Predictability	2.10	0.01	1	0.147
Foundation ✻ Emotion	2.16	-0.01	1	0.166

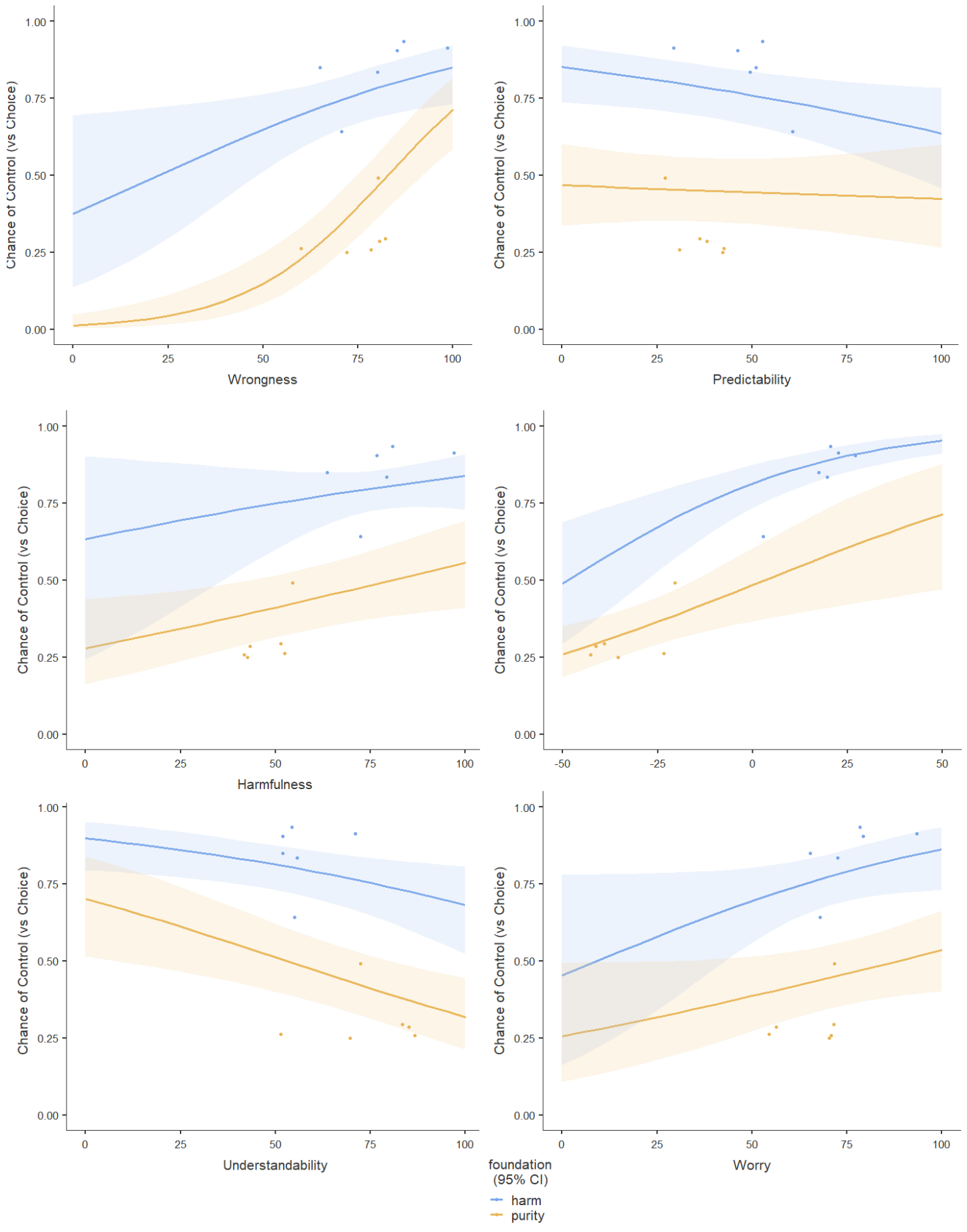
Finally, to assess whether the set of judgments can fully account for the Partner Management differences between Harm and Purity violations, three models were compared and their predictive performance was assessed via their Bayes Factors. The three models were: the Full Model which included all judgments and

moral foundation, the Judgments Model which contained only the judgments and the Foundations model which contained only which Moral Foundation was violated. The Full Model performed decisively better than the Judgments Model, $BF_{10} = 6.54 \times 10^{107}$ and the Judgments Model performed decisively better than Foundation-Only Model 9.55×10^{461}. These results suggest that there are differences between Harm and Purity violations besides the judgments we probed in order to account for the different Partner Management reactions.

As with Experiment 1, to determine whether the Moral Foundation is decisive in predicting whether people engage in Partner Choice or Partner Control we plotted model predictions from the data when the terms for “Purity” and “Harm” are switched and everything else is kept the same. Results confirmed that switching and Harm and Purity predominantly switches the predicted reactions from Partner Choice to Partner Control and vice versa.

Figure 2.12.

Judgments predicting Partner Management



Note: Points represent empirical averages per vignette

Discussion

Purity violations once again resulted in different Partner Management reactions compared to Harm violations. As before, Harm violations were more associated with Partner Control and Purity violations were more associated with Partner Choice. This was true even if Partner Management would be conducted by a third party rather than the participants imagining they would have to punish the perpetrator themselves (although punishment favorability is marginally reduced, Harm violations still lead to Partner Control and Purity violations to Partner Choice).

Once again, Purity violations incurred different kinds of judgments than Harm violations. We assessed how participants felt about the moral wrongness of the act, how harmful they believe it to be, the predictability of the violator, how easy it is to understand their motivations and how worried they would be around the perpetrator, including after they were sufficiently punished. Harm violations were viewed as slightly more morally wrong than Purity violations but it should be noted that, unlike the first study, participants leaned more liberal (undergraduate student sample) here so this cannot be taken as a fully representative sample of the population. Purity violators were seen as less predictable and participants found their motivations more difficult to understand.

Surprisingly, the idea that punishment would be less efficacious for Purity violations was not supported here as participants felt worried that both Purity and Harm violators would cause future harm and importantly, to the same extent, felt this would be true even after the perpetrator was sufficiently punished. In other words, perceived punishment efficacy appears to be the same for purity and harm violations using this measure.

In addition to these judgments, we also assessed the different emotional responses that these violations elicited. In line with previous work, Harm violations were more associated with anger while Purity violations were more associated with disgust. In addition, as predicted, when considering having committed the violation themselves, Purity violations were more associated with shame and Harm violations were more associated with guilt.

We also assessed how these judgments and emotional reactions are associated with different Partner Management reactions. As predicted, when motivation is more difficult to understand and the perpetrator is perceived to be less predictable, reactions were more likely to be Partner Choice, whereas more harmful acts were most likely to result in Partner Control. In addition, as predicted, the more disgusting or shame-inducing an act was, the more likely it would result in Partner Choice.

Despite these judgments and emotional reactions explaining some of the variance in Partner Management between Harm and Purity, once again, it is clear that there are other fundamental differences between these violations, as the moral foundation being violated was still a decisive factor in predicting Partner Management even when accounting for the probed judgments and emotional reactions.

General Discussion

Partner Management Across Violations and Judgments

Many perceived moral violations can trigger anger or disgust, often driving us to pursue retribution or instead to simply avoid the violator, potentially dissolving whatever social relationship we may have had with them. This research has looked at what features of an action lead us to these different kinds of responses, finding that Purity violations reliably lead to Partner Choice and Harm violations to Partner

Control. This was demonstrated across two experiments with different response scales and samples and this finding also persisted across political and demographic lines.

When encountering a social partner who has caused harm, we make act-based judgments about blame and punishment and employ Partner Control to punish the perpetrators for causing harm to a victim. In contrast, for cases of moral impurity, we instead rely more often on Partner Choice behaviours (e.g. avoidance, deselection and social group ostracization) to avoid agents we deem to be morally bad or defective or perhaps simply not like "us" in some important way.

The BIS and BAS, (the regulatory systems associated with avoiding and approaching stimuli) have previously been established as a key factor in regulating our own moral behaviour (Janoff-Bulman and Carnes 2013; Sheikh and Janoff-Bulman 2010; Janoff-Bulman et al. 2009) (e.g. by regulating feelings of shame and guilt). As predicted, shame and guilt and their 3rd party mirrors of disgust and anger, respectively, predicted Partner Management reactions. Purity violations led to feelings of disgust and Partner Choice (avoidance of the perpetrator) and when imagining being the perpetrator themselves, participants were more likely to feel shame. Harm violations were more likely to lead to anger and would cause participants to seek retribution by approaching the wrong-doers. The role of these systems in regulating behaviour towards social partners should be explored in more detail in future work as the BIS and BAS vary in important ways for moral judgments such as the concreteness and strictness of a norm. Further work should focus on how this relates to Partner Management.

When Purity violations were particularly associated with strong character-based judgments, participants were most likely to prefer Partner Choice. These judgments include diagnosticity of moral character, understandability and discomfort around the perpetrator, their predictability and the understandability of their motivation. The findings regarding predictability echo other recent research on predictability and moral judgement (Walker et al. 2020). Predictable social partners facilitate cooperation even with individuals who may have harmed another in the past. More work should focus on the perceived predictability of social partners and how this affects Partner Choice.

Individual Differences

Conservatives and Liberals care about different moral foundations (Feinberg & Willer, 2019; Graham et al., 2009; Hatemi et al., 2019; Kivikangas et al., 2020). This finding is well established and our results corroborate this (Experiment 1). However, we also intended to ask a different question: Given that conservatives purportedly care more about Purity violations than liberals, do they react differently to the perpetrators of those violations? Remarkably, the answer appears to be no. Not only do liberals react (rather than doing nothing) at the same rate as conservatives to Purity violations, they also react in precisely the same ways, that is, by overwhelmingly preferring Partner Choice to Partner Control. Indeed, this remained true across all of our demographic measures despite the fact that higher Relational Mobility was associated with stronger moral judgments.

Conservatives are more likely to be part of smaller and tighter communities and social groups (Maxwell, 2019; Waytz et al., 2019) which should be especially evident when assessing the rural-urban political divide. Therefore, we explored the theory that in smaller, tighter groups, such as rural communities, a potentially unsafe social partner has a greater chance of harming the particular person or the people close to them, so it will be more beneficial to be sensitive to signals of moral character (like purity violations) that would indicate the potential for future harm. However, we found that rural and urban participants showed no difference in judgments. While this is evidence that our hypothesis is incorrect,

this might instead be the result of using self-report measures for urban and rural location as it has been argued that these measures may be inadequate to capture the relevant information (Nemerever and Rogers 2021). To support this, we point to the fact that our rural participants (based on self-report) were, surprisingly, not more likely to be politically conservative. Therefore, future work should explore if and how differences in rural and urban ecology may explain these effects and we propose using more objective measures for future work as recommended by Nemerever and Rogers (2021) and potentially using a different sampling procedure.

One drawback to the way in which these experiments have focused on Partner Control is that punishment was largely framed as interpersonal. We hypothesised that fewer opportunities for Partner Choice would lead to a greater reliance on Partner Control but there is a good reason this might not be the case, specifically, for interpersonal punishment. People may well avoid punishing others even when there is less opportunity for Partner Choice because of the danger of cycles of retaliation since one is less likely to be able to avoid the person they have punished. Therefore, it is possible that the relationship between punishment and Partner Choice availability might be more complicated. Although we looked at the differences between 1st and 3rd party Partner Control (experiment 2) we only measured individual differences in experiment 1 since the second experiment focussed on an undergraduate subject pool and so would not generate enough variability in the demographic and individual differences measure we used. It is possible stronger effects would be found when focusing more on third-party and institutional punishment or punishment organised around dominance hierarchies. Each of these provides useful Partner Control machinery that avoids retaliatory aggression in tighter groups. Therefore we suggest future work should look at Relational Mobility, Partner Choice availability and the role of 3rd party and institutional punishment and suggest that Partner Control may be more relied on through these means when Partner Choice is less available.

One might question why Purity violations are punished at all. One possibility is that for very tight groups, Partner Choice is less available, leaving only Partner Control to enforce norms. If this is the case then measures of Relational Mobility (the freedom and opportunity societies offer individuals to choose and dispose of interpersonal relationships) (Awad et al. 2020; Thomson et al. 2018) may predict the propensity towards Partner Control over Partner Choice. However, we found evidence against this hypothesis. It is possible that using the Relational Mobility scale in this way is problematic as it asks participants about other people around them rather than themselves and has predominantly been used to assess sociological and cultural differences between groups rather than differences between individuals. Therefore, we suggest more directly measuring individuals' opportunity for Partner Choice and measuring the environmental influences on this to understand how this may affect Partner Management.

Monism vs Pluralism

Early research has focused on perpetrators causing intentional harm to a victim, (Kohlberg 1973; Piaget 1932). This focus on only harm has been criticised by moral foundations theorists (Graham et al. 2009; Graham et al. 2013) but the results of this study might be taken as evidence for a somewhat monist view of moral judgement that centres around harm. Monist accounts, isolating harm, exist in the literature, c.f. Dyadic Morality (Schein and Gray 2018; Gray et al. 2014; Schein and Gray 2015). These use a more wide-ranging definition of harm that includes so-called "spiritual harm", the Theory of Dyadic Morality reiterates the idea that harm (caused intentionally by a moral agent to a moral patient) underlies all forms of moral judgement, (Schein and Gray 2018). However, since these accounts attempt to redefine harm, they have been met with criticism on the grounds that the definitions become too broad, lack explanatory

power and contain a model of "harm pluralism" that lacks parsimony (Graham et al. 2018). Instead, we offer an alternative account that requires no re-definition of harm but rather argues that moral judgments can either centre on an actual harm to a victim or they can centre on the potential for future harm to potential future victims where people perceive the possibility of future harm from moral agents who have displayed defective moral character (Chakroff 2015). Therefore, it is possible to argue for a harm-based morality where 1. agents punish harm perpetrators (Partner Control) and 2. largely avoid, exclude or ostracise, potential perpetrators of future harm who have not yet caused harm (Partner Choice). For example, agents that appear to lack the inbuilt disgust mechanisms usually associated with the domain of moral purity (eg. incest, bestiality etc) found in normal social partners. Despite this, explicit judgments related to future harm did not drive these differences. In one sense, this is unsurprising because if the sensitivity to moral impurity is to avoid future harm, this mechanism does not need to involve explicit judgments about future harm. Rather they could just be driven by affective states (Haidt 2001) (disgust, discomfort), as even for other social primates (where that kind of explicit cognition may not be possible), Partner Choice drives cooperation (Schino and Aureli 2017). In other words, it is possible that a sensitivity to those committing Purity violations helps one avoid social partners who would otherwise cause future harm and therefore it is selected for without the need for explicit judgments about future harm.

On the other hand, our results may instead be interpreted as evidence for moral pluralism. We assessed many kinds of judgement and even emotional responses. Even when including all these factors in models predicting Partner Management, knowing what individual foundation was violated between harm and purity was still strongly predictive, suggesting there are more fundamental differences between these foundations that have not been captured. There still appears to be something special about Purity violations compared to Harm violations that are not accounted for by these factors.

These results indicate that a wide array of different judgments and affective states moderate the strength of a selected Partner Management strategy. However, it is not yet clear what is decisive in which strategy is preferable and therefore what it is about Purity and Harm that drive their respective differences in Partner Management.

Future Directions

Findings from Moral Foundations theorists have highlighted a fascinating puzzle; why are some actions viewed as morally wrong when there is no apparent objective harm and why is this largely true only for a subset of individuals? One proposed solution is that our social preferences and moral inclinations are shaped by evolutionary selection pressure from temporal and geographical variations in parasite stress, (Fincher and Thornhill 2012; Thornhill and Fincher 2014; Tybur et al. 2016). Cross-cultural differences in contact with disease and ecological hazard are said to drive a "behavioural immune system" that will vary in its sensitivity to perceived "impure" acts. However, rather than a direct danger from contamination, we propose that more general indications of defective moral character can result in different reactions compared to when there are consequential harms, and potentially, the features of social groups could dictate which Partner Management strategy will be the most adaptive. When actions appear to have caused no objective harm, our aversion to them will stem from the signals the actions send about the moral character of the agents themselves, where judgments are chiefly concerned with who someone is, rather than just what they have done. This will depend on whether purity violations are genuinely perceived to have no victim. Therefore, future work should isolate purity violations that contain a wronged victim to see if such violations are more likely to result in Partner Control. In addition, if conservative religious communities are more likely to punish purity violations, this might be explained by

the inclusion of a *wronged* party, in this case, the wronged party in question is a deity who is wronged by one failing to observe its rules. This should be explored in more detail in future work.

Finally, This work has focused only on the foundations of harm and purity. It remains an open question how people prefer to react to violators of the remaining foundations of Authority, Loyalty and Fairness. Therefore, further work should focus on Partner Management reactions to violations of these other foundations. In addition, this work has focused only on negative Partner Management (punishment and avoidance) but has not explored the positive dimensions of Partner Choice (selection) and Partner Control (direct reciprocity). That is, what kinds of reactions result in direct reward compared to Partner Choice (e.g. increased inclusion in social activities, friendships etc).

Conclusion

Moral Foundations Theorists argue that purity violations are not at all about harm (Graham et al., 2011, 2018), whereas those who argue for the Theory of Dyadic Morality (harm-centric morality) argue that purity is simply a form of "spiritual harm" (Gray et al., 2021; Schein & Gray, 2018). We adopted the position that foundations of harm and purity both involve harm, where harm violations involve a specific harm to a victim and purity violations give character-based signals that indicate the potential for future harm from that particular perpetrator. This position generates the hypothesis that Purity violators would be treated differently to Harm violators, namely that the former would be avoided to mitigate future harm whereas the latter would be punished and we found strong support for this hypothesis. This is important because it can help us understand the mechanisms that drive social group ostracization compared to what drives direct punishment.

Differing sensitivity to moral foundations across the political divide has been used to explain many different phenomena in political psychology. We have demonstrated that violations of different foundations result in fundamentally different reactions but remarkably those reactions remain the same across the political spectrum, the rural-urban divide and across different levels of Relational Mobility. Understanding how people react to different kinds of moral violation and why they react that way is important both for policymakers and those wishing to understand the deeper structures and purposes behind moral judgement. Understanding how we manage our social partners is integral in comprehending how social hierarchies form, how we structure our institutions and may help answer more fundamental questions about social group cohesion.

This paper presents a first step in understanding how we manage social partners in response to different kinds of moral violation and also explores the different kinds of judgement that are most associated with each strategy. There are potentially many aspects to a moral violation that would drive preferences for one strategy over the other. We hope to explore what other aspects drive people's preferences for each strategy in future research.

Finally, we hope these findings may contribute to the developing functional account of moral purity sensitivity. We maintain that viewing morality through the lens of Social Partner Management can provide further insight into our moral minds.

Chapter 3. Does Counterfactual Requirement Explain the Side-Effect Effect?

Abstract

The Side-Effect-Effect describes the strange phenomenon that people ascribe intentionality to bad outcomes compared to good ones. This effect has largely resisted explanation but a recent model of intentionality ascription provides a promising avenue for exploration. This model states that if a side-effect is counterfactually required for the desired outcome then we say that the side-effect was intentional. This means we ascribe intentionality to a side-effect if, in counterfactual worlds where the side-effect would not happen, the desired outcome would not be obtained. Although this model has been successful in accounting for data regarding judgments of particular trolley dilemmas, we test this account using the classic vignettes from the Side-Effect-Effect literature. First, we test whether assumptions about counterfactual requirement can account for the Side-Effect-Effect (Study 1) and second we test the extent to which counterfactual requirement intervenes on intentionality and moral judgments when it is systematically manipulated and made explicit (Study 2). In both studies, we find that counterfactual requirement has no effect on intentionality judgments. Therefore, at first glance, it appears that Counterfactual Requirement is unable to explain or even intervene on Side-Effect Effect cases. However, we discuss whether alternative explanations (including those that maintain the importance of counterfactual reasoning) are potential candidates to explain both sets of data.

Introduction

How do we know when to hold others responsible for the outcomes of their actions? When should we blame or praise them and when should draw inferences about their character based on these outcomes? To answer these questions, our common-sense explanations for the outcomes of an agent's action distinguishes between the mere consequences of the action and the agent's genuine intentions. This distinction is crucial for social reasoning, moral judgement, teaching and learning and extends to legal policy, and economics. But if we are to trust our folk judgments in making this distinction we must understand the cognitive processes that underlie those judgments. It is argued that intentionality has to do with inferences over the state of mind of the actor, such as their desires and beliefs (Malle and Knobe 1997). In other words, all that is required is whether or not you think the agent desired that outcome and believed that their action would obtain it. However, despite this initial straightforward outline, it appears that the inner workings of folk intentionality are far more elusive than first expected, especially when evaluating the side-effects of an action.

When probing the intuitions of participants on whether or not an action was intentional, rather than simply reflecting inferences about the actor's state of mind, responses appear to be sensitive to the valence of outcomes (Knobe 2003) and normative considerations (Hindriks 2014; Knobe 2010). This finding has been termed the Side-Effect Effect (SEE) (also known as the Knobe Effect). It emerges when people consider two contrasting cases, one good and one bad, where people appear to make asymmetrical ascriptions of intentionality. The classic case asks people to consider a CEO who is about to adopt a new policy that will make lots of money but will also harm the environment. The CEO, who is indifferent regarding the environment, institutes the policy and the company makes lots of money. Participants are asked whether the CEO harmed the environment intentionally and most (82%) say *yes*. However, if the

policy helps rather than harms the environment (everything else is kept the same) most (77%) say *no*, the CEO did not intentionally help the environment.

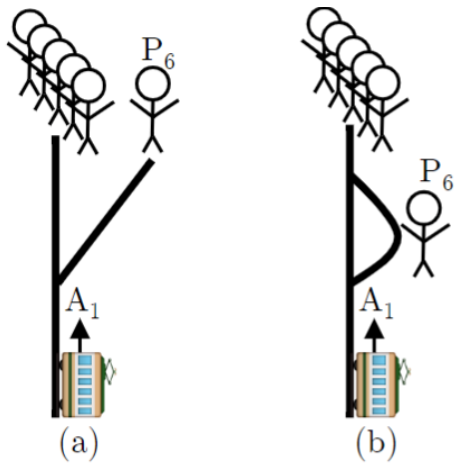
There are many such cases like this and many competing explanations, for example the harm case is more diagnostic of character than the help case and there is more justification for blame and many other explanations (e.g. Nakamura 2018; Dalbauer and Hergovich 2013; Lin et al. 2019; Hindriks et al. 2016; Laurent et al. 2020; Laurent et al. 2019; Cova 2016). In addition, these effects have been replicated in other languages and cross-culturally (Burra and Knobe 2006; Cova and Naar 2012) and with children (Leslie et al. 2006; Pellizzoni et al. 2009) and beyond the individual to group-intentionality (Michael and Szigeti 2019). Finally, these asymmetries extend beyond intentionality to judgments of freedom (Phillips and Knobe 2018) and knowledge (Paprzycka-Hausman 2018; Dalbauer and Hergovich 2013). Empirical data gathered from participants' intentionality judgments across these cases have been remarkably difficult to understand. Many models aimed at predicting some subset of the data by incorporating other factors besides the agent's state of mind have also largely failed when applied to all the data, see Cova (2016) for a review.

Rather than attempting to resolve this debate, which is beyond the scope of this paper, we instead focus on a new and promising area of enquiry around the SEE. The aspect we focus on is counterfactual reasoning, where one samples from possible alternative states of affairs that could have occurred rather than what actually occurred. In other words, one reasons about what might have been. This tendency is widely known to be integral to many psychological processes, for example, causal reasoning (e.g. Dablander 2020; Kominsky and Phillips 2019; Morris et al. 2019; Willemsen and Kirfel 2019; Lagnado et al. 2013; Quillien 2020; Icard et al. 2017), moral judgement (e.g. Byrne 2017; Kleiman-Weiner et al. 2015; Migliore et al. 2014; Bernhard et al. 2021; Phillips et al. 2015), and action selection (e.g. Icard et al. 2018; Phillips et al. 2019; Phillips and Cushman 2017; Phillips and Knobe 2018).

Importantly, counterfactual reasoning has also been implicated in attributions of intentionality (Halpern and Kleiman-Weiner 2018; Kleiman-Weiner et al. 2015; Quillien and German 2021). Kleiman-Weiner et al., (2015) propose a computational account of intentionality that centres around the concept of *counterfactual requirement*. When judging the intentionality of a side-effect, this model states that if the side-effect was counterfactually required for an outcome, then that side effect was intentional. This means that we assign intentionality to a side-effect if, in counterfactual worlds where the side-effect would not be present, there would be no reason to pursue the desired outcome using the present means. To illustrate this, two famous trolley dilemmas are described. In both cases, suppose that a trolley is heading towards five people and will kill them unless it is diverted to an adjacent track, where there is a single bystander who will be killed if the trolley is diverted. In the first case (figure 3.1a), diverting to the second track will always save the five people, whether or not there is a bystander on the adjacent track. However, in the second case (figure 3.1b), the adjacent track loops around connecting back to the first track so if there is no bystander for the trolley to hit then the trolley won't stop and the five people will still be killed. Therefore, in the second case hitting the bystander is required to stop the trolley but in the first case, it matters not for the desired outcome whether the bystander is present. In the first case, there is reason to divert the trolley either way, but in the second case, the (act-consequential) optimal policy *counterfactually depends* on there being a bystander to stop the trolley. Importantly, when testing this model on the described trolley cases, side-effects that were counterfactually required were much more likely to be judged as intentional (Kleiman-Weiner et al. 2015) lending substantial support for the model. (See also the distinction between "in order *to*" and "in order *that*" [Knobe 2010]).

Figure 3.1

Classic and Loop variant of the trolley dillma (Kleiman-Weiner et al. 2015)



How might this relate to the SEE? Counterfactual Requirement may drive the SEE because of the assumptions people make across the two cases. People would imagine environmental harm happens because the company is "extracting value" from it. If the environment turns out not to have been harmed, this would necessarily mean that the company failed to make money. Participants would be assuming environmental harm is counterfactually required. No such assumption would be made in help cases, participants would expect that if the environment was unaffected the company would still succeed in making money from the policy. In other words, people would be more likely to expect that harming the environment was integral to the policy succeeding in making money but that helping the environment was not.

This model was not originally proposed to explain the SEE and participants may not make these assumptions, nevertheless, it is also possible that making them explicit (i.e. whether or not the help/harm is counterfactually required) may supervene on intentionality judgments, moderating or even reversing the SEE. Therefore, whether or not it is explanatory, it may still affect intentionality ascriptions in these cases in important ways that bear on other current theories. For example, Counterfactual Requirement may only influence one of the cases, (either harm or help). If this occurs then it would provide support for accounts that favour the Interpretive Diversity Hypothesis (IDH) (Laurent et al. 2020; Laurent et al. 2019; Nichols and Ulatowski 2007; Cova ; Cushman and Mele 2008). The IDH is intended to explain SEE by claiming that intentionality has different definitions depending on the context and so asymmetries in response represent different processes simply because participants are implicitly answering different questions. For example, they may take intentional to mean: “desired”, or “was in control”, or “did so willingly”, depending on the context of the question. If differences in counterfactual requirement only affect certain cases, it may be due to the specific concept of intentionality that is activated by that case.

Finally, the effect of counterfactual requirement on intentionality has only been tested in trolley type cases and may not affect classic SEE cases at all. If this is true then the proposed model for the intentionality of side effects may be insufficient as it would fail to account for a wide array of data, or the role of counterfactual requirement may be more complicated and may influence a third variable differentially for the different cases.

Across two studies (n = 395) we test the relevance of counterfactual requirement in the classic CEO case, finding that the model fails to account for these classic judgments but also, surprisingly, that manipulating counterfactual requirement (the integral part of the model) also has no effect on intentionality judgments in these cases. We consider what this means for both simple and more complex models based on counterfactual reasoning in the General Discussion.

Experiment 1 - Testing Asymmetrical Intuitions for Counterfactual Requirement

This study probes participants' assumptions about the classic CEO case, assessing whether they believe harming or helping the environment was integral (counterfactually required) or simply tangential for making money. If counterfactual requirement has any explanatory power over the classic SEE then it predicts that participants would be more likely to find that harm to the environment is integral and, in contrast, help to the environment would be more likely seen as tangential in successfully earning money from the policy.

Participants

We recruited 200 participants through Prolific Academic (aged between 18 and 62, M_{age} = 26.8, SD_{age} = 8.58, 57.1% male). The questions for this study were attached to the end of a separate study (pre-registration https://osf.io/bjxpc/) but the randomisation procedure is not affected by that study. The pre-registration for this study can be found here:https://osf.io/xryh8/.

Procedure

Participants were split into one of two independent groups where they read modified versions of either the classic help or harm side-effect cases. The relevant change is that the potential side-effect fails to come about. The vignette states *"*A CEO of a company is sitting in his office when his Vice President of R&D comes in and says *'We are thinking of starting a new programme. We forecast that it could allow us to increase our profits. However, it could also [harm/help] the environment.'* The CEO responds that he doesn't care either way about the environment as long as the policy is profitable. The programme was carried out but the environment was unaffected.*"*

Participants responded to two questions regarding the vignette. The first question asked on a 5-point scale *"Even though the environment wasn't affected, how do you think this policy affected profits?"*. Responses ranged from *"made much more than predicted*" to *"made much less than predicted"* The second question asked with a yes/no response *"Do you think the policy would have made more money if the environment was [helped/harmed]?"*

Results

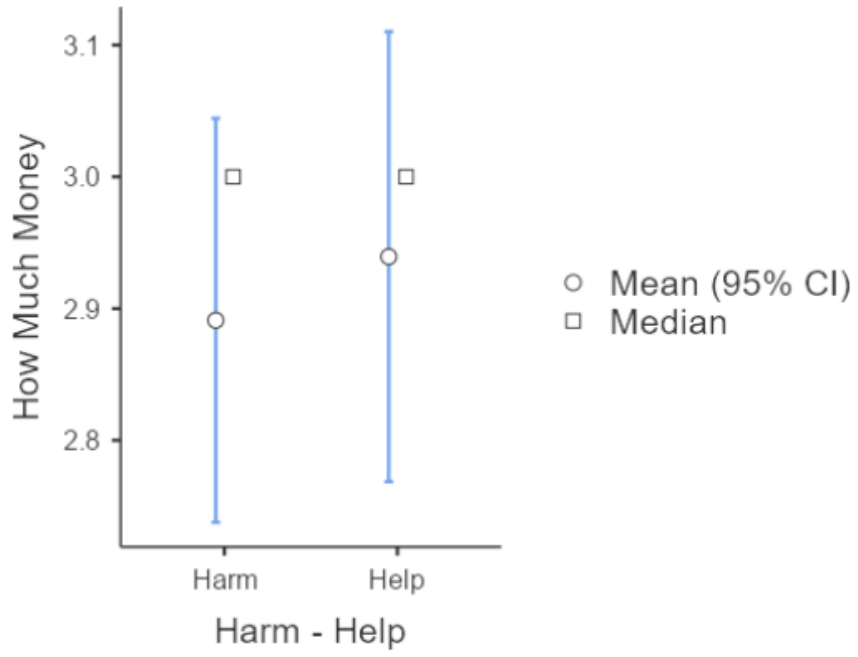
All analyses were performed in Jamvoi (Şahin and Aybek 2019).

An independent samples Welch's t-test was conducted to compare judgments about how profits were affected between the harm and help groups. There was no significant difference between the groups $t(195.31) = 0.41$, $p = 0.680$, $d = 0.06$, $BF_{10} = 0.167$, Bayesian analysis indicates evidence against this effect. See Figure 3.2.

A chi-square test of independence was performed to assess whether the harm and help group differed in their response to whether the company would have made more money if the environment was affected. The difference was non-significant, χ^2 (1, N = 200) = 3.36 , p = 0.067. Bayesian contingency tables suggest strong evidence against an effect $BF_{10} = 0.02$. 41/101. 41% said the Company would have otherwise made money in the harm case and 54% said the company would have otherwise made money in the help case.

Figure 3.2.

Comparing harm and help groups on of predicted effect of profits if the environment was not affected



Discussion

Surprisingly, there was no significant difference between the help and harm cases on either question. This indicates that participants do not make different assumptions about whether counterfactual requirement is present between the two cases and therefore the counterfactual requirement model cannot explain the difference between harm and help cases in judgments of intentionality.

Experiment 2 - Manipulating Counterfactual Requirement

Although this model is unable to explain the SEE, it was never intended to do so. However, the model would still predict differences in intentionality when the presence or absence of counterfactual requirement is made explicit. In this second study, we directly manipulate whether or not the harm and help to the environment is counterfactually required to make money in order to see whether this affects judgments of intentionality. We also probe moral judgments about blame and moral character.

Participants

We recruited 195 participants recruited through Prolific Academic (aged between 18 and 62, M_{age} = 25.5, SD_{age} = 7.91, 62% male). The questions for this study were attached to the end of a separate study (pre-registration: https://osf.io/utsm9/) but the randomisation procedure is not affected by that study. The pre-registration for this study can be found here: https://osf.io/g8euq/.

Procedure

Participants are split into one of four independent groups where they read modified versions of the classic vignettes. Two variables are manipulated, the valence of the outcome (help vs harm) and the presence or absence of counterfactual requirement ("only make money if" vs "will make money either way"). The vignettes were as follows, *"There is a new policy that could make a company lots of money. In some circumstances, this policy may harm the environment [and will only make money if the environment is [harmed/helped]/ [but it will make money either way (i.e. whether it [harms/helps] the environment will have no effect on profits)]. The CEO of this company does not care either way about the environment and only cares about making money. The CEO implements the policy and the environment is [harmed/helped]."*

Participants respond to five questions regarding this vignette, *"Did the CEO [help/harm] the environment intentionally?"* yes/no response; *"How blameworthy/praiseworthy was this act?"*, on a 5-point scale from extremely blameworthy to extremely praiseworthy; and *"How good or bad would you rate the moral character of the CEO?"* on a 5-point scale from extremely bad to extremely good.

Results

All frequentist analyses were performed in Jamvoi using default settings unless specified otherwise (Şahin and Aybek 2019) and Bayesian models (quantifying the levels of evidence for and against each predictor/model) were fitted using the BayesFactor package with default settings (Morey et al. 2015). The bayesTestR package was used for calculating Bayes Factors of inclusion (comparing all possible models with the term against all equivalent models with that term removed) (Makowski et al. 2019).

A generalised linear model predicting intentionality was calculated with Help/Harm and Required as predictors. The model was significant, χ^2 (3) = 106.83, p < .001, explaining 40% (McFadden's R^2) of the variance. Harm cases were significantly more likely to be rated as intentional compared to help cases (B = -2.97 , 95% CI [-4.00, -1.94], OR= 0.5, 95%CI [0.02, 0.14, p < .001, BF_{10} = 2.38 x 10^{16}). This replicated the classic SEE. However, Counterfactual Requirement did not significantly predict intentionality (B =1.04, 95% CI [-0.08, 2.17], OR = 2.84, 95% CI [0.92, 8.79], p = 0.070, BF_{10} = 0.087), nor did it interact with Help/Harm (B = -1.10, 95% CI [-2.73, 0.53], OR = 0.33, 95% CI [0.07, 1.69], p = 0.185, BF_{10} = 0.204) (Bayesian evidence was against these effects), see figure 3.3 A.

A linear model predicting Blame-Praise was calculated with Counterfactual Requirement and Help/Harm as predictors. Results indicated that the model was significant, F(3, 191) = 45.25, p < .001, R^2 = 0.42). Harm cases were significantly more blameworthy than help cases (B =1.48, 95% CI [1.13, 1.83], t = 8.41, p < .001, BF_{10} = 6.42). However, once again Counterfactual Requirement was not a significant predictor , (B = -1.8, 95% CI [-0.53, 0.17], t = -1.02, p = 0.307, BF_{10} = 0.599) (Bayesian test was insensitive), nor its interaction with Help/Harm (B =-0.07, 95% CI [-0.57, 0.43], t = -0.27 , p = 0.785, BF_{10} = 0.210) (Bayesian evidence was against this effect). See figure 3.3 B.

Finally, a linear model predicting Moral Character was calculated with Counterfactual Requirement and Help/Harm as predictors. Results indicated that the model was significant, F(3, 191) = 24.98, p < .001, R^2 = 0.28). CEOs in Harm cases were rated as having significantly worse moral character than in help cases (B =0.96, 95% CI [0.68, 1.24], t = 6.66, p < .001, BF_{10} = 1.69 x 10^{12}). However, once again Counterfactual Requirement was not a significant predictor , (B =-0.02, 95% CI [-0.3, 0.26], t = -0.14, p = 0.890, BF_{10} = 0.242) (Bayesian Evidence against this effect), nor its interaction with Help/Harm (B =-0.16, 95% CI [-057, 0.25], t = -0.77, , p = 0.440, BF_{10} = 0.269) (Bayesian evidence was against this effect). See Figure 3.3 C.

To determine how character, blame and intentionality are associated with each other a correlation matrix was calculated for just harm cases and just help cases, see Figure 3.3. For harm cases, intentionality was not significantly associated with blame-praise (p = 0.147, Pearson's r = 0.146) but was significantly associated with character (p < 0.001, Pearson's r = 0.336). Blame-praise and character were significantly associated (p = 0.003, Pearson's r = 0.298). For harm cases, intentionality was not significantly associated with blame-praise (p = 0.118, Pearson's r = -0.166) nor with character (p = 0.081, Pearson's r = -0.180). Blame-praise and character were significantly associated (p < 0.001, Pearson's r = 0.393),

Figure 3. 3

Models Predicting Intentionality, Blameworthiness & Moral Character

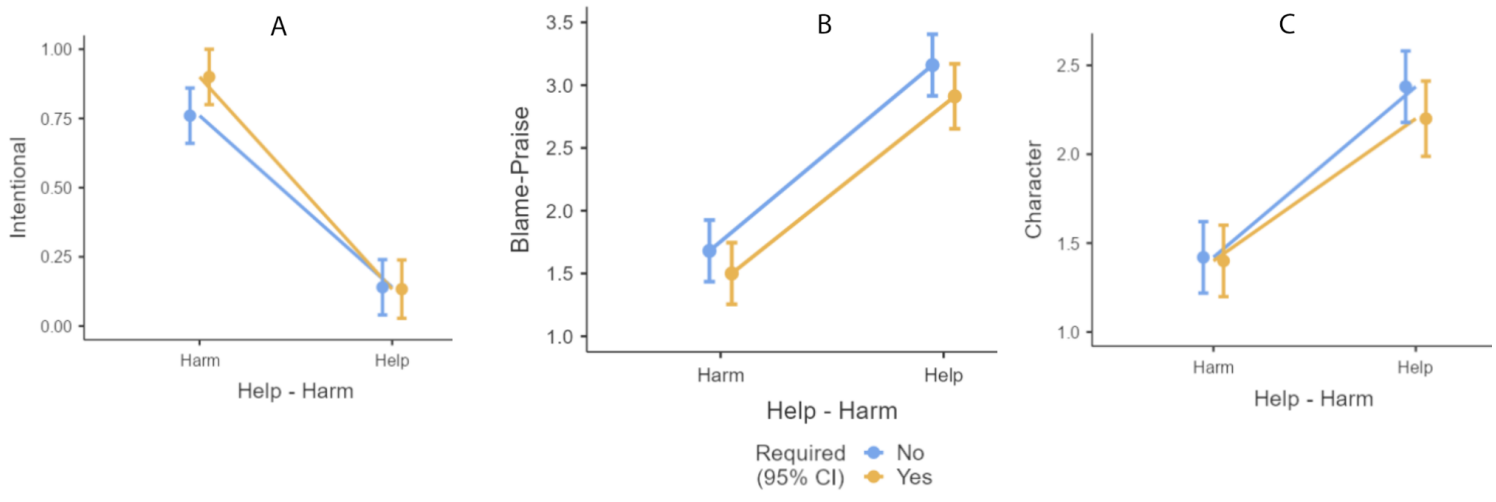
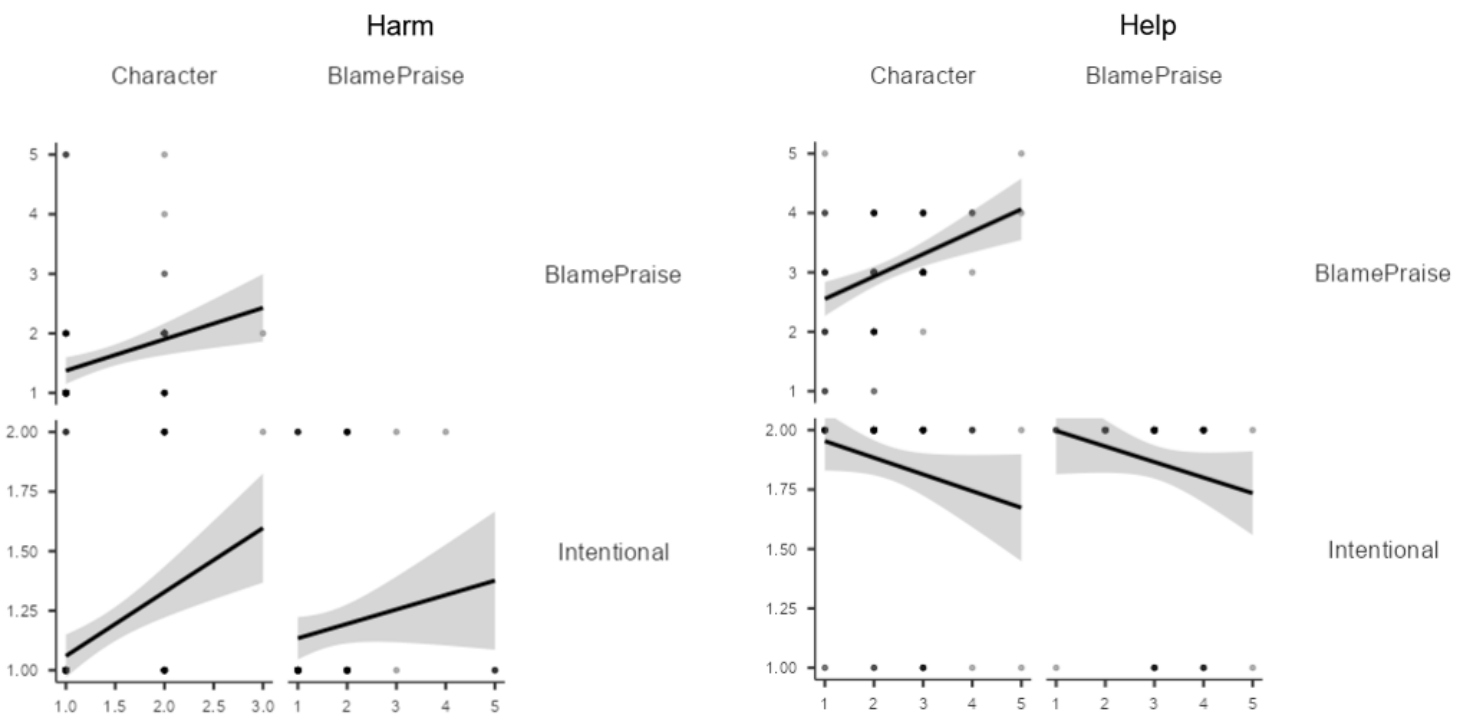


Figure 3.4

Correlation Matrices between Intentionality, Blameworthiness & Moral Character



Discussion

Surprisingly, there was no significant effect of counterfactual requirement on intentionality judgments, nor on the secondary moral judgments (blame/praise and moral character). These results show, unequivocally, that for cases where the classic SEE is apparent, counterfactual requirement appears to have no significant effect and Bayesian analyses indicate evidence against this effect. Either the counterfactual requirement model only applies to trolley type cases or there is a third variable that counterfactual requirement intervenes on in the trolley cases but not in the CEO cases. These possibilities are discussed further in the General Discussion.

General Discussion

When we make certain judgments we often imagine how things might have been rather than how they actually turned out. This tendency is called counterfactual reasoning. One specific consideration we make when considering counterfactual worlds is whether the same outcome would come about if certain elements are removed. For example, after a burglary one might wonder whether if they had left their lights on the burglar may have bypassed their house. In other words, they consider the extent to which

having the light off was counterfactually required for the burglary to happen. The idea of counterfactual requirement has been proposed as a fundamental aspect for predicting and explaining how people ascribe intentionality to the outcomes of an action, especially when those outcomes are side-effects of the intended outcome (Halpern and Kleiman-Weiner 2018; Kleiman-Weiner et al. 2015). This model centres on the idea that if a side-effect is counterfactually required for the desired outcome to be obtained, then those side-effects will be deemed to be intentional. For example, if a single bystander is sacrificed to save a group of people, then people will consider worlds where the bystander wasn't present to determine whether the bystander's death should be thought to be intentional. The model has been successful in predicting responses to precisely these kinds of trolley dilemmas (Kleiman-Weiner et al. 2015) and proves to be a useful piece of the puzzle for understanding how we make moral judgments about blame and moral responsibility (Halpern and Kleiman-Weiner 2018). Given its success, we tested it on cases that characteristically exhibit the Side-Effect Effect, where bad actions are seen to be more intentional than good actions.

First, we tested whether this model can explain, partially or fully, this effect by probing whether people make different assumptions about counterfactual requirement between cases with good side effects (helping the environment) and cases with bad side effects (harming the environment). However, it does not appear that these assumptions are made and therefore it is unlikely that this model has any explanatory power for the classic SEE. Second, by modifying the cases, we tested how intentionality judgments are affected by counterfactual requirement, by noting its presence or absence explicitly in the vignettes. Surprisingly, this had no discernible effect on either intentionality judgments nor on judgments of blame, praise and moral character.

Why might there be a discrepancy in the model's predictive power between the trolley dilemmas and the SEE cases? Perhaps one explanation is that the goal (saving people) in the sacrificial dilemma is seen to be morally good, whereas the CEO's goal is making money, which is at most morally neutral or morally bad. It has been previously shown that evaluations over the desired outcome (rather than just the side-effects) make a difference in the SEE (Waleszczyński et al. 2019). However, there is not yet any reason to assume that counterfactual requirement only matters when the goal is morally good. Another option is that in the looped track trolley dilemma (Figure 3.1b), where the bystander is used as a means rather than an end, killing the bystander becomes an explicit goal rather than simply a side effect. In other words, counterfactual requirement simply delineates between actual side-effects and goals. However, the data here do not support that interpretation because, in CEO-help cases, where helping the environment is counterfactually required, participants still judge the action to be unintentional.

Alternatively, it is possible that these data sets, taken together, lend support to the Interpretive Diversity Hypothesis, (Laurent et al. 2020; Laurent et al. 2019; Nichols and Ulatowski 2007; Cova ; Cushman and Mele 2008). The IDH states that there are multiple concepts for which the word "intentional" refers.

For example, it is argued that in some cases people consider whether an act was done "intentionally" but in other cases (morally bad cases) one instead asks whether the act was "done with foreknowledge" (Nichols and Ulatowski 2007; Laurent et al. 2020; Laurent et al. 2019). Alternatively, it has been proposed that there are "two and half folk concepts" of intentionality where one examines desires or beliefs and some intervention of moral judgments (Cushman and Mele 2008). It may be that counterfactual requirement is of special importance for some interpretations of "intentional" but has no bearing on others. However, it is difficult to see what concept of "intentional" is activated in the trolley cases but is not present in either CEO-harm or CEO-help cases as none of the proposed candidates currently fit the

bill. This is because desires, beliefs and foreknowledge do not meaningfully differ in the trolley cases in a way that would be consequential to these accounts.

One final option, which is more parsimonious than the previously discussed candidates, draws from recent insights in causal cognition (eg. Morris et al. 2019; Willemsen and Kirfel 2019; Lagnado et al. 2013; Icard et al. 2017; Quillien 2020). In this model, Quillien & German (2021) seek to repurpose the causalist model of intentionality (Davidson 1963; Davidson and Davidson 1980) as a psychological model.

This model states that outcomes are intentional if the agent's desires and beliefs *cause* the outcome. Importantly, causal models of cognition also centre on the importance of counterfactual requirement. Stated simply, rather than the side-effect being counterfactually required for the desired outcome, this model states that the deciding factor is whether the *agent's attitudes and beliefs* towards the outcome were counterfactually required for the outcome to occur. This is true for both intended outcomes and side-effects. Notably, Quillien and German (2021) show that even for the classic CEO case, judgments about whether the agent's attitude towards the side-effect *caused* the side effect almost perfectly correlates with judgments of intentionality. The reason why this is the case, it is argued, is that counterfactual reasoning is biassed by norms and expectations in the same way that is apparent in the SEE, a point also made by Bernhard et al. (2021).

How might this model, or similar models, explain the two datasets? In the trolley cases, counterfactual requirement changes the attitude towards the bystander; they are now a means rather than an end. This is important because using someone as a means in this way is perceived to be counter-normative (Kleiman-Weiner et al. 2015), even for young children (Levine and Leslie 2020). Counterfactually relevant alternatives that are sampled in counter-normative cases are biassed towards cases where the agent instead acts normatively (Kominsky and Phillips 2019; Quillien 2020; Quillien and German 2021). This would suggest that in the looped track case, people intentionally kill the bystander *because* they believe they can use the bystander as a means. This attitude is counter-normative and so people imagine counterfactuals where this attitude isn't present and find that the lever isn't pulled so the bystander survives. This leads to judgement that the attitude towards the bystander was causal and therefore killing the bystander was intentional. In contrast, for the CEO who harms the environment, either as a means or simply as a potential foreseen side-effect, is always acting counter-normatively. Our data supports this because judgments of blame and negative moral character persist for harm cases irrespective of the presence or absence of counterfactual requirement. Therefore the causalist models may account for both the data from trolley cases and our data from classic SEE cases. However, more direct tests of this are recommended, especially because there are many other vignettes where the SEE emerges.

Conclusion

Intentionality is integral to social cognition in general and moral cognition in particular as it drives, for example, judgments about moral responsibility and blame. However, it appears as if morally salient cases are judged in a different way to morally neutral cases, and the strangeness of intentionality judgments even extends beyond that. Much work has been done trying to gain insight into why this happens but we have focused on one specific insight, that counterfactual reasoning may underpin how we ascribe intentionality.

However, when testing a previously successful counterfactual reasoning model of intentionality on cases where this strangeness arises, the model is no longer able to predict the data. Rather than abandoning this model wholesale, we argue that counterfactual requirement may be important specifically as it pertains to

causal reasoning and in agreement with recent models, it is possible that previously problematic normative causalist accounts may hold the key when repurposed as a psychological model that incorporates recent insights into causal reasoning.

Chapter 4. Suffering and Dying: How Speciesism Matters for Assessing Extreme Harms

Abstract

Previous work shows that Speciesism appears to account for our preference for the welfare of humans over non-human animals, over and above the mental capacities we ascribe to humans compared to other non-human animals. However, intuitions regarding moral standing become more complex when we compare judgments regarding animal or human suffering and judgments regarding animal or human killing. Across four studies (n = 671) we explore the different drivers of these judgments for different animal and human targets. When comparing torture and killing, perceived Agency best predicted which act is considered worse. When considering torture or killing in isolation, perceived Experience was instead the determining factor. The view that humans are special over and above their mental capacities seems to drive judgments regarding killing, but when making judgments about suffering it appears to play little or no role. This finding may help explain differing attitudes to voluntary human euthanasia compared to animal euthanasia and why animal rights activists may focus on the suffering rather than the deaths of animals.

Introduction

Suppose that a violent sadist has decided to capture a deer so that he may torture it for his own amusement. His aim is to inflict as much suffering as possible over several days before finally letting the deer go. In contrast, consider that this same sadist wishes to simply kill the deer, again, purely for his own amusement. He does so painlessly by shooting the deer through the head. Which of these acts do you feel is more morally wrong, the deliberate infliction of suffering or the painless taking of the deer's life? Now consider instead that the sadist's target is a human being. Either the sadist wishes to torure him for a short period or he wishes to kill him by shooting him through the head. Do you feel differently about which kind of harm is more morally wrong?

Killing and torture are both extreme harms, but which is worse seems to depend on whether the victim is a human or animal. Comparisons between killing and suffering become all the more concrete when, for example, drafting policy on assisted suicide or discussing the morality of animal euthanasia. Assisted suicide is illegal in many countries across the world and is forbidden by most religious teachings despite the fact that such prohibitions have the potential for a great deal of suffering. We also have a fundamental aversion to the loss of human life, described as one of our "Sacred Values" where it would be deemed outrageous to trade human life, under any circumstances as if it were an economic good (Tetlock 2003). For animal lives it's different. Animal lives under factory farming are often viewed as economic goods and the mercy killing of a suffering animal is often considered to be the morally appropriate act. Indeed suffering animals are put down routinely by veterinary practitioners. Causing death and causing suffering appear to be fundamentally different kinds of harm and they appear to be treated differently depending on the moral patient.

Intentional harms have been the focus of much work in moral psychology. Moral Foundations Theory offers a pluralistic account, arguing that harm is one of five or more foundations and is integral to most moral judgement (Graham et al. 2018; Graham et al. 2011; Haidt and Joseph 2008; Graham et al. 2009). Others have argued that harm, properly defined, is rather the cornerstone of all moral judgement (Schein et al. 2016; Schein and Gray 2015; Gray et al. 2014; Schein and Gray 2018). Although we wish to reduce harm, rather than using simple utilitarian cost-benefit analysis, we are extremely reluctant to endorse causing harm to another, even if it is purportedly for the greater good (Greene 2016; Cushman et al. 2010; Greene 2014; Greene 1294; Mikhail 2007). However, this constraint appears to be more relaxed when it comes to sacrificial dilemmas involving animals both in hypothetical dilemmas, (Caviola et al. 2020) and in laboratory experiments where participants can interact with animals (Bostyn et al. 2018). This work shows that people are reluctant to sacrifice a person to save five others and they become all the more reluctant if this must be done directly and in close proximity. However, when given the option to shock a rat to save five other rats from being shocked people are far more willing to do this. It is clear that we have a distinct and explicit sensitivity to intentional harm but these experiments do not involve direct comparisons between causing suffering and taking a life, especially across different agents.

History paints an unpleasant picture of how humans have treated animals in our recent past. The philosopher Descartes famously viewed animals as mindless robots; mere automata who lack the ability to neither think nor feel (Descartes 1989; Miller 2013). This has been historically used to justify many acts which we would now consider barbaric, for example, vivisection without the use of anaesthesia (Allen and Trestman 2017). The practice of cat burning, for both entertainment and for superstitious reasons pervaded mediaeval Europe up until the 1800s (Darnton 2009). Even now, animal rights campaigners and activists increasingly draw our attention to the suffering of animals under conditions of factory farming (Singer 1995) leading to the widespread adoption of veganism (Gruen and Jones 2015) or reductitarianism/flexitarianism (Derbyshire 2016) (reduced meat consumption). In addition, prior to the 1990s in the US, there was no official consensus on the existence of animal pain (in terms of phenomenological experience) and so veterinary practitioners were trained to ignore the signs of pain and thus withhold pain relief (Rollin 1989).

Despite this, animal companionship has existed throughout most of human history (Kean and Howell 2018) and even Descartes' contemporary, Henry More, wrote to him condemning his views as "deadly and murderous" (Goodman et al. 2012). Public opposition towards animal experimentation continues to grow (Goodman et al. 2012) and our understanding and recognition of animal consciousness has seen remarkable progress, evolving into mature fields encompassing animal cognition, neuroscience and practical ethics (Allen and Trestman 2017). Given our complex history recognising the moral worth of animals, what psychological processes underpin our moral concern and how might these processes differentiate between suffering and killing?

Philosophers have argued over what faculty an entity has to have in order to be worthy of moral consideration. In the original formulations of Utilitarianism, the concept of utility was centred on positive and negative mental states like happiness and pain (Bentham 1789; Mill 1887; Sidgwick 1981). Jeremey Betham famously declared "the question is not, can they reason, nor can they talk, but, can they suffer?" (Bentham 1789). In contrast, Preference Utilitarianians focus on the agent's ability to have preferences and the complexity of those preferences,(Singer 2011; Hare 1981; Singer 1995) for example, the preference to continue living rather than their feelings of happiness or suffering. The fact that humans have much larger and more complex preferences compared to animals is integral to why we believe human life to be of particular importance (Singer 2011). Singer (2011) also highlighted the moral difference between killing

a being who experiences pain but is not able to meaningfully reflect about themselves, compared to killing a more intelligent being. Kantians require the entity to be "rational being" where the agent's intelligence and ability to think is the focus Kant (1796/2002).

Gray and colleagues (2007) find that the kinds of minds people automatically attribute to different agents are seen to vary across the two dimensions of agency and experience. Agency relates to the ability to reason and act on the world (eg. memory, planning, self-control) whereas experience relates to the ability to feel (eg. pleasure, pain, desire). These findings have proved to be particularly useful, revealing many important aspects of our moral psychology including moral typecasting (Gray and Wegner 2009), moral responsibility in artificial intelligence (Bigman et al. 2019), religious conceptions of suffering (Gray and Wegner 2010) and dyadic completion (Gray and Wegner 2011). Importantly, animals across different species are seen to vary greatly in their capacity for agency and experience (Gray et al. 2007).

The idea that moral concern may rest on perceived agency reflects Kant's view on Personhood and the idea of a "rational being" (Kant and Schneewind 2002). Intelligent agents such as humans are said to be deserving of rights and the protection of their interests particularly due to their ability to reflect on the reasons for their action. The idea that moral concern, instead, rests on perceived experience more closely resembles Hedonistic Utilitarianism and Bentham's focus on the capacity to suffer (Bentham 1789). For a review of the different ways one can assess levels of phenomenal consciousness and how this relates to matters of moral patienthood see (Muehlhauser 2018).

When considering how important it is to mitigate the suffering of a specific animal, one might imagine its capacity for feeling would be the obvious criteria that influence judgement. However, consider cases of painless death (e.g. euthanasia or idealised farming). Here, it is less clear why the capacity to suffer would be the key decider for this because, other things being equal, the animal would not suffer. The aversion to killing, compared to aversion to the suffering of agents may be treated very differently, especially between animals and humans. On preference-based models of utilitarianism, one recognises that moral patients can have preferences, both in continuing their life and also over their future well-being and the ability to reflect on these preferences is rooted in intelligence (Hogg 2005). For these reasons, it is possible the perceived agency of the patient may play a stronger role, specifically in judgments of killing, and perceived experience may play a stronger role for judgments about suffering.

Alternatively, there may be more fundamental reasons why human life or human suffering would be treated differently. The idea that humans are in some sense exceptional and that human welfare must always categorically take precedence over the suffering of animals is called *speciesism* (i.e. moral worth determined only on the basis of species membership) (Horta 2010; Caviola et al. 2019; Caviola et al. 2020; Fjellstrom 2002). On the one hand, we may recognise in humans some unique features salient to moral concern (Kagan 2016; Grau 2016). The ostensibly unique human capacities proposed include abstract thinking, social and familial ties, emotion and language, but evidence for many of these features have been firmly established in the animal kingdom and they differ a great deal between the species (Gruen 2017). Instead, we may hold prejudicial or biassed attitudes purely on the basis of species membership, often likened to racial prejudice or sexism (Fjellstrom 2002; Caviola et al. 2019). In fact, Speciesism better predicts moral worth judgments over and above the role of mind perception alone (perceived agency and experience) (Caviola et al. 2019). This is to say that the value people place on humans compared to other animals is not fully accounted for by differences in intelligence and sentience that people believe exists between the species. Caviola and colleagues (2019) argue that Speciesism has the same ideological roots as other prejudices and find these measures to be highly correlated.

People also place extra deontological constraints for humans, compared to animals, in sacrificial dilemmas (Caviola et al. 2020). That is, participants considered it more permissible to harm animals if it meant a larger number of animals would be saved but the same was not true for judgments involving harm to humans. This follows the idea that ethics consists of utilitarianism for animals but Kantianism for people (Nozick 1974). Similarly, it is argued animals do not have rights to not be killed such that it is permissible to sacrifice a chicken to save five others (Thompson 1990).

To explain how we differentiate between agents in terms of suffering and killing, we have presented many different possibilities. It is possible that humans are seen as special both for judgments regarding suffering and judgments regarding killing or it is possible that humans are only seen as special for judgments regarding one of these. If the latter is true this might be explained by the mental capacities we ascribe to humans compared to animals. For example, if humans are perceived to have higher agency and agency is the decisive factor when making judgments about killing, then humans will be judged to be special particularly for judgments of killing. Another possibility is that there is little difference in judgments between non-human animals and therefore simply whether or not the agent is human will be the decisive factor. Finally, it is possible that this is true only for judgments of killing and not suffering, where speciesism is focused on the sanctity of human *life* rather than being more broadly about human *welfare*.

Across four studies (n = 671) we explore these different possibilities. Study 1 has participants compare an unspecified hypothetical animal to a hypothetical human being as the target of a sadist who either wishes to torture or kill them. Study 2 replicates this procedure across a series of different animal and human targets while measuring the perceived Agency and Experience of each target. Study 3 replicates this between subjects and Study 4 has participants make judgments about either suffering or killing in isolation. Sample sizes for these studies were chosen based on previous research on mind perception and animal moral worth judgments

Experiment 1: Torture vs Killing in Animals and Humans

We test the intuition that when deciding which extreme harm is worse out of torture and killing, judgments differ when the moral patient is a human compared to an animal. This study was pre-registered and can be found here: https://osf.io/kvwnz/

Participants

We recruited 150 participants from the United States and Canada (aged between 19 and 73, M_{age} = 35.45, SD_{age} = 10.38, 61% male) from Amazon Mechanical Turk in exchange for a small payment of $2.40. The questions for this study were attached to the end of a separate study which can be found here: https://osf.io/m5jrn/. Participants were recruited based on the criteria that they must have had at least 50 previously accepted HITs and a prior approval rating above 90%.

Procedure

To test whether an asymmetry in judgements over extreme harms exists between animals and humans, participants see vignettes about someone who either tortures or kills for fun using a between-subjects design where participants are asked about either humans or animals. The vignettes are as follows: "*Imagine there is a person, who for fun targets random [animals/people] and kills them instantly by shooting them in the head. Now imagine there is a person, who for fun targets random [animals/people], captures them and for a*

short period and tortures them to inflict pain before letting them go. In general, out of the act of killing an [animal/human being] or torturing a [animal/human being], which do you feel is morally worse?"

Participants judge whether the act of torture or the act of killing is worse on a 7-point scale ranging from (1) *"torture is much worse*" to (7) *"killing is much worse*".

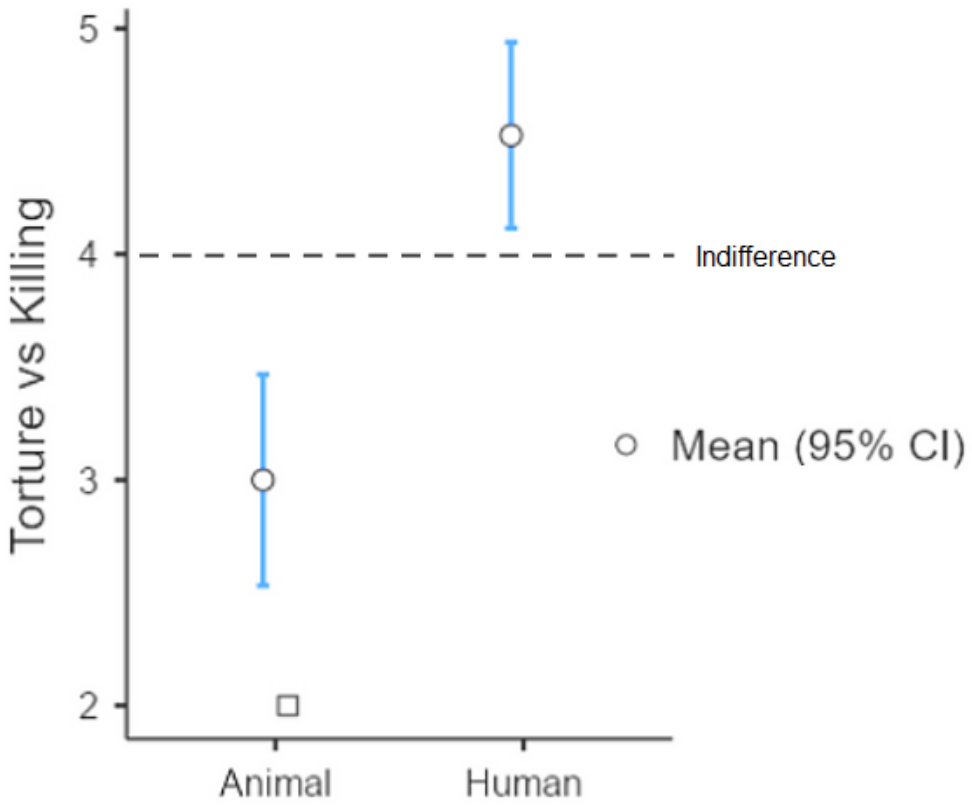
Results

To test whether judgments from each group differed from indifference (4 on the 7-point scale) we preregistered two one-sample T-tests. Results indicated that for both animals and humans, killing and torture judgments were significantly different from indifference, for animals (M = 3, SD = 2.07), t (74) = -4.19, $p < 0.001$, $BF_{10} = 523$ and for humans (M = 4.53, SD = 1.81), t (73) = 2.51, $p = 0.007$, $BF_{10} = 4.7$. As predicted, for humans people judged killing as being worse than torture and for animals, the torture was worse than killing, see Figure 4. 1.

Furthermore, an independent samples, Welch's T-Test, revealed that judgments for animals and humans significantly differed from each other t(147) = 4.8, $p < 0.001$, Cohen's d = 0.79, $BF_{10} = 4039$.

Figure 4. 1.

Comparing the Moral Wrongness of Torture to Murder for Animals and Humans



Discussion

Study 1 confirmed that people make asymmetric judgments about extreme harms between animals and humans. For animals, torturing is judged to be worse than killing but for humans, the reverse pattern is observed. The experiments that follow test explanatory hypotheses that may explain this asymmetry.

Experiment 2: Torturing vs Killing Across Different Patients

It is possible that this asymmetry exists when comparing any animal-human pairing such that this difference can be explained only by appealing to humans as special in some way compared to other animals (Speciesism). Alternatively, judgments may differ even between different animals and different humans based on the attributes of each target. One important attribute is the perceived minds of the targets, which is the focus of this study. Participants are given a selection of different humans and

animals, to test whether these judgments reflect a general difference between animals and humans or specific differences between each target. We measure how participants perceive the minds of each target and test whether this predicts differences in judgment of torture vs killing. This study was pre-registered and can be found here: https://osf.io/kvwnz/

Participants

We recruited 161 participants (aged between 18 and 28, M_{age} = 18.93, SD_{age} = 1.15, 10% male) from Warwick University's psychology undergraduate student subject pool. Participants received course credit for taking part in the experiment.

Procedure

To test whether differences in judgement exist only between humans and animals or whether differences exist between all targets, we use a within-subjects design where participants make judgments across 22 target patients (11 humans and 11 animals). The animals are: octopus, crab, family dog, cow, sheep, chicken, orangutan, spider, ant, pigeon and police dog. The humans are: doctor, CEO, teacher, child, baby, athlete, warehouse worker, person with a severe mental disability, you - the participant, accountant and bus driver.

To test whether the perceived mind of the target accounts for judgments regarding torture vs killing, mind perception is measured using 2 five-point scales one for perceived agency: "*How much each of these targets is capable of thinking. That is, how much is each able to plan ahead, reason about ideas, think, and have general intelligence.";* and one for perceived experience "*How much each of these targets is capable of feeling. That is, how much can they have emotions and feelings, and experience sensations of pain and pleasure.*

For measuring judgments of torture vs killing, participants are told: "*For each agent please imagine: First that someone, for fun, kills them instantly and painlessly. Second that someone, for fun, captures them and for a short period painfully tortures them before letting them go."* Participants are asked which act is morally worse on a five-point scale between torture and killing (0 = torture is much worse, 5 = killing is much worse). For exploratory purposes, we also code animals based on whether or not they are farm animals.

Results

Pre-registered, one-sample T-tests show judgments for all patients significantly differ from indifference, all $p < 0.001$ (holm corrected), where for animals torture was worse and for humans killing was worse.

All frequentist mixed models initially specified all random slopes for all predictors and then random slopes were removed one by one to find the maximal converging model, which is recommended as best practice for linear mixed-effects models (Barr et al. 2013). Frequentist analyses were performed using Jamovi v1.6 (Şahin and Aybek 2019), mixed models were calculated using the GAMLj module for Jamovi. Bayesian models (quantifying the levels of evidence for and against each predictor/model) were fitted using the BayesFactor package with default settings (Morey et al. 2015). All the Bayesian mixed models had random slopes specified for all predictors. The bayesTestR package was used for calculating Bayes Factors of inclusion (comparing all possible models with the term against all equivalent models with that term removed) (Makowski et al. 2019).

A linear mixed-model regression was performed to predict torture vs killing judgments with Patient-Type (animal or human) perceived Agency and perceived Experience as predictors. (Frequentist: random intercepts and slopes for Patient-Type by Participant ID),

The results indicate that Patient-Type, Agency and Experience were significant predictors of torture vs killing judgments. Patient-Type: $F(1, 246.36) = 10.96$, $B = 0.37$, $SE = 0.11$, $p = 0.001$, $BF_{10} = 35.79$; Agency: $F(1, 3386.2) = 60.01$, $B = 0.19$ $SE = 0.02$, $p < 0.001$, $BF_{10} = 5.13 \times 10^{11}$; Experience: $F(1,3483.8) = 12.04$, $B = -0.13$, $SE = 0.04$, $p < 0.001$, $BF_{10} = 20.4$. There were also significant interactions between Patient-Type and Agency such that Agency had a stronger effect for Humans than Animals: $F(1,3310.04) = 5.27$, $p = 0.022$, $BF_{10} = 0.55$; Patient-Type and Experience such that Experience had a stronger effect for Humans than Animals: $F(1,3385.53) = 12.85$, $p < 0.001$, $BF_{10} = 1.82$.

Taken together, these analyses indicate that there is evidence for differences between animals and humans and evidence that perceived agency is a predictor of judgments between targets. Although there is evidence that perceived experience is also a predictor of these judgments, the Bayesian analysis suggests that this study was not sensitive enough to draw strong conclusions from this.

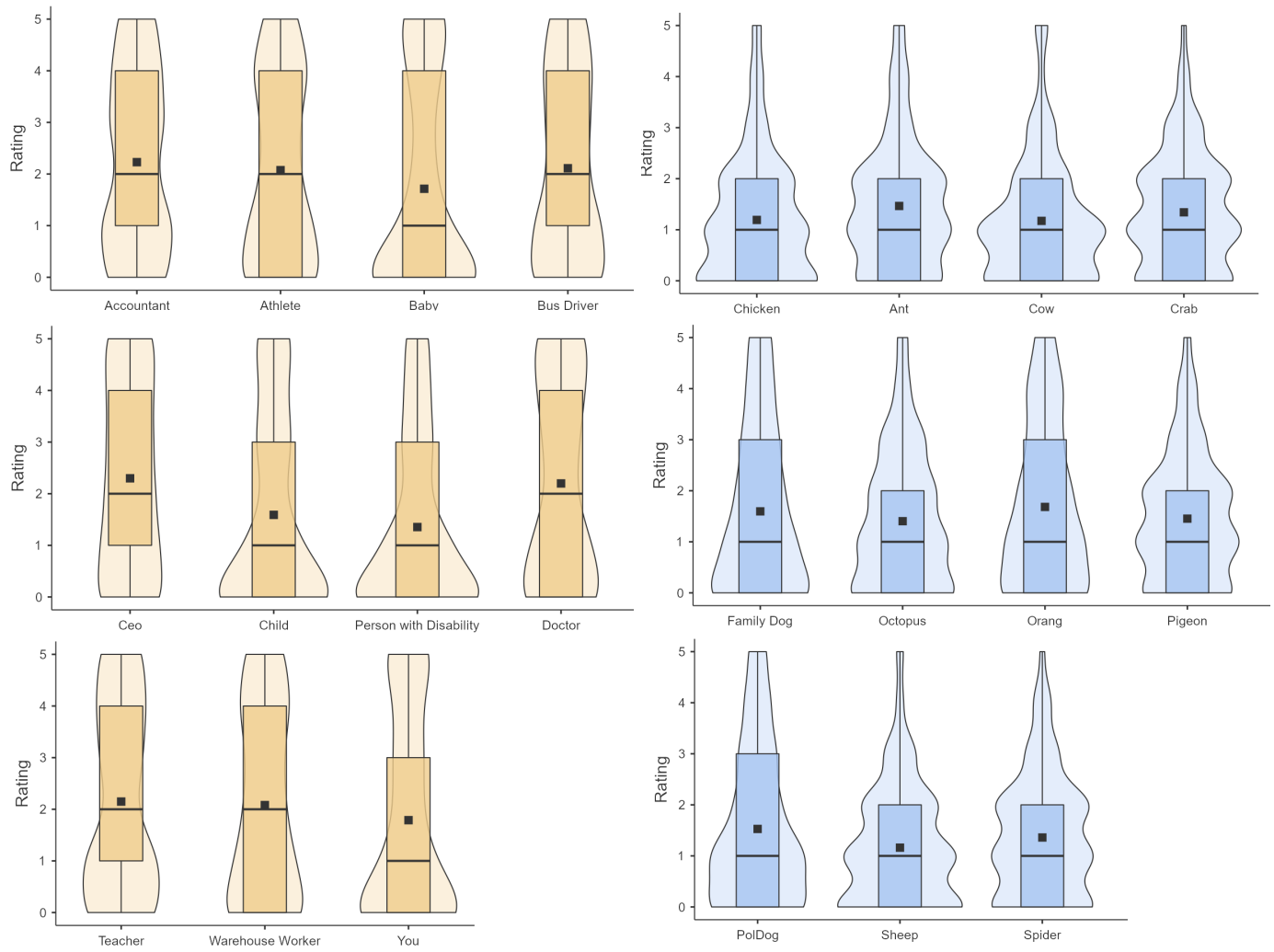
In order to determine whether speciesism, mind perception or both best explain the variance in judgments of torture vs killing, three Bayesian models were compared. The speciesism model contained only Patient-Type as a predictor; the mind perception model contained only Agency and Experience; the full model contains all three predictors. The full model substantially outperformed both the speciesism model $BF_{10} = 1.196 \times 10^{11}$ and the mind perception model $BF_{10} = 12.00$, suggesting that both mind perception and speciesism play a role in judgments comparing extreme harms.

Importantly, despite finding differences between animals and humans, the mean for all patients showed torture was rated as worse than killing, see Figure 4.2 and Figure 4.3. However, in line with study 1, the distributions appear to indicate that people were much more likely to say that killing is worse than torture for humans compared to animals.

In order to statistically test whether this was the case, responses were re-coded as either “killing is worse” (3 - 5) or “torture is worse (0 - 2), see Figure 4.4. A generalized mixed-model regression was performed to predict recorded torture vs killing judgments with Patient-Type (animal or human) as a predictor with random slopes by Participant ID (maximal converging model). Results confirm that participants were significantly more likely to choose Killing for humans than for animals $\chi^2(1) = 13.71$, $B = 0.37$, $SE = 0.11$, $p < 0.001$.

Figure 4. 2

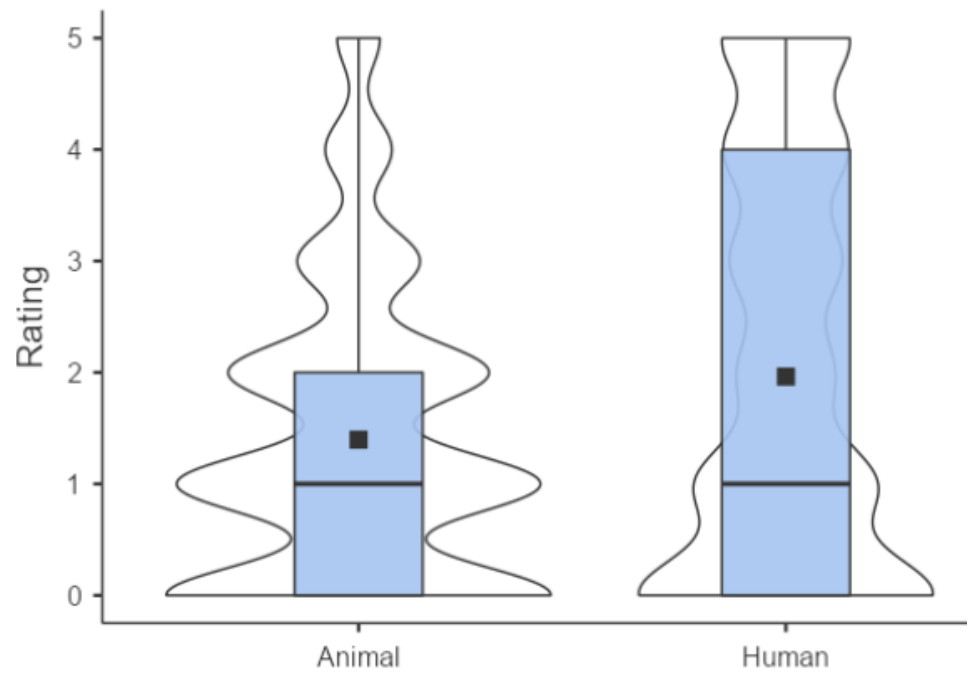
Distributions of Torture vs Murder Responses per Patient



Note. *< 2.5 indicates torture is worse, > 2.5 indicates killing is worse*

Figure 4. 3

Distributions of Torture vs Murder Responses for Animals and Humans

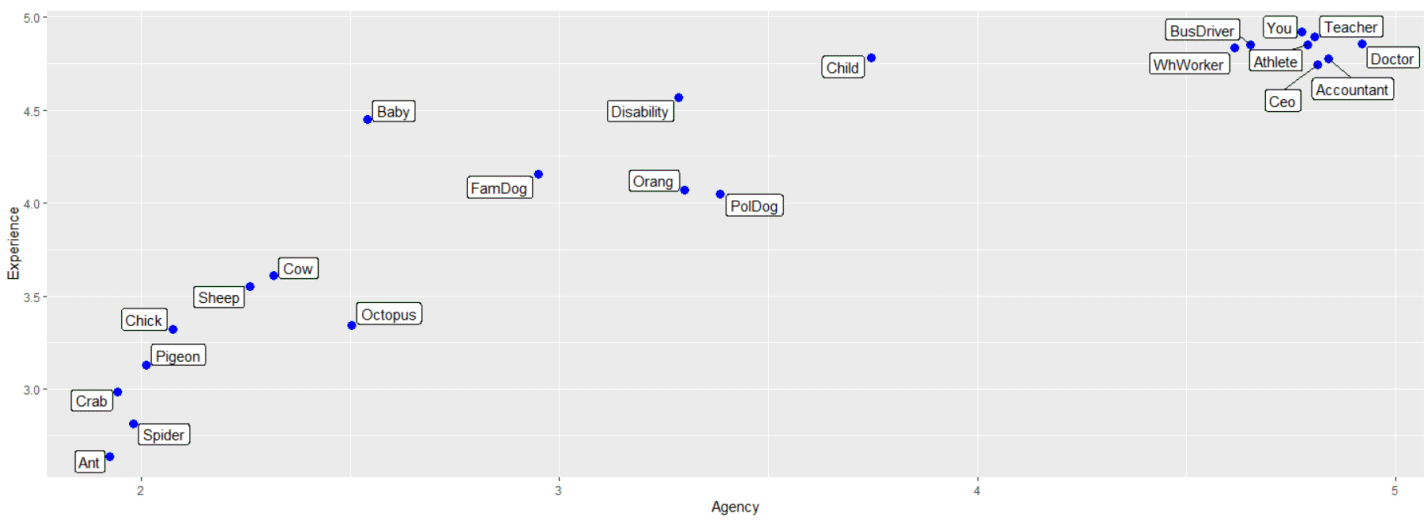


Note. *< 2.5 indicates torture is worse, > 2.5 indicates killing is worse*

To visualise how the targets differed in terms of mind perception, Agency and Experience were plotted for each target, see Figure 4.4

Figure 4.4

Minder Perception for each patient



It is possible that we would get particularly strong effects for farm animals simply because we are so used to them being killed for food. Therefore, we coded whether or not the target was a farm animal or not and looked at the data only for animal targets to perform exploratory analysis. Linear mixed-model regressions were fitted predicting torture vs killing with Farmed vs Non-farmed, Agency and Experience as predictors, all main effects are specified as having random slopes. Whether the animal was a farm animal or not significantly predicted judgments $F(1, 188.78) = 17.22$, $B = -0.28$, $SE = 0.07$, $p < 0.001$, $BF_{10} = 29.26$. However, all other predictors are non-significant in this model, Agency, $F(1, 209) = 1.11$, $B = 0.055$, $SE = 0.052$, $p = 0.293$, $BF_{10} = 8.30$; Experience, $F(1, 105.78) = 0.063$, $B = -0.011$, $SE = 0.042$, $p = 0.803$, $BF_{10} = 0.17$; Experience:Agency, $F(1, 675.91) = 0.43$, $B = 0.023$, $SE = 0.035$, $p = 0.50$, $BF_{10} = 0.11$; Experience:Farm, $F(1, 316.00) = 0.314$, $B = 0.034$, $SE = 0.060$, $p = 0.576$, $BF_{10} = 0.08$; Farm:Agency = $F(1, 326.16) = 1.71$, $B = -0.093$, $SE = 0.071$, $p = 0.19$, $BF_{10} = 0.35$. In addition, adding Farm to the full model (Mind Perception and Patient-Type) substantively improves the predictive performance $BF_{10} = 5.47$

These results indicate that judgments differ between farmed and non-farmed animals and these results are best explained by how we perceive the minds of farmed animals, specifically their perceived agency. To explore this further, linear mixed models predicting a perceived agency were fitted with farmed vs non-farmed as predictors with random intercepts for participant and random slopes by participant ID. Farm animals are seen to have significantly less agency than other animals $F(1, 1609) = 41.92$, $B = 0.28$, $SE = 0.04$, $p < 0.001$, see Figure 4.5 & 4.6.

Figure 4.5

Perceived Agency of Farmed and Non-Farmed Animals

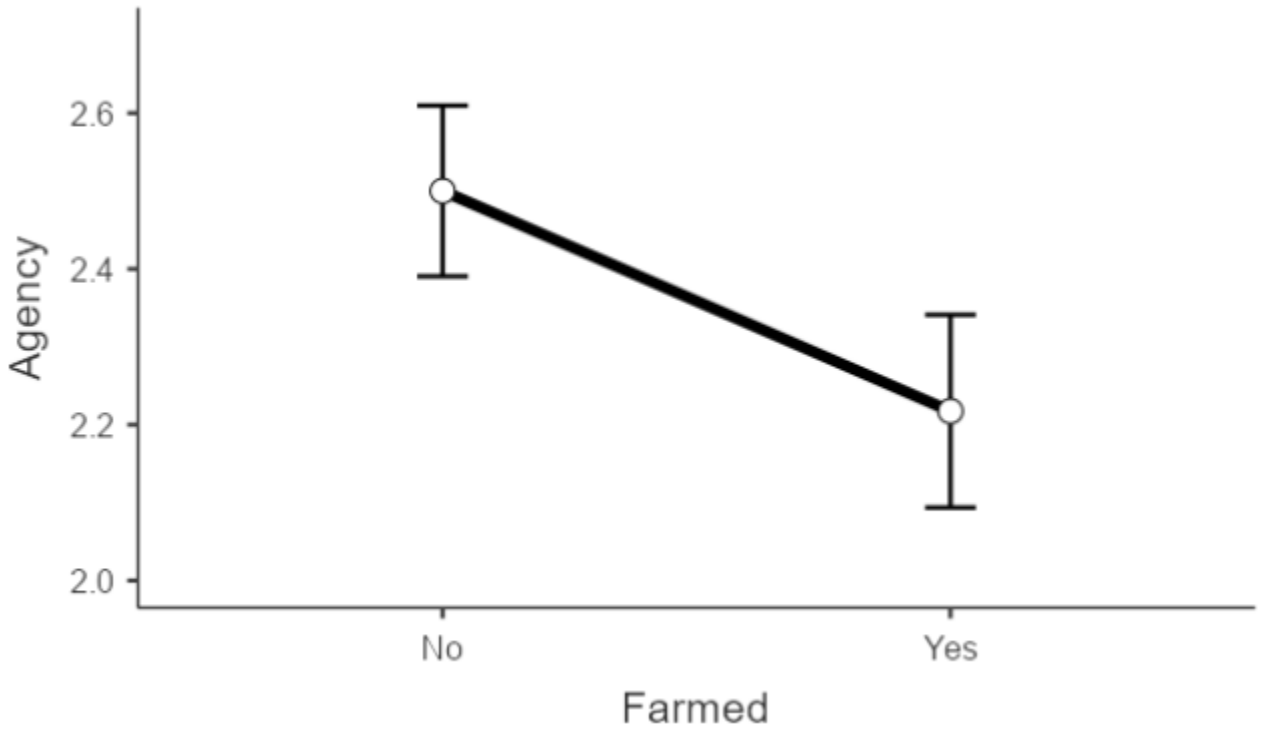
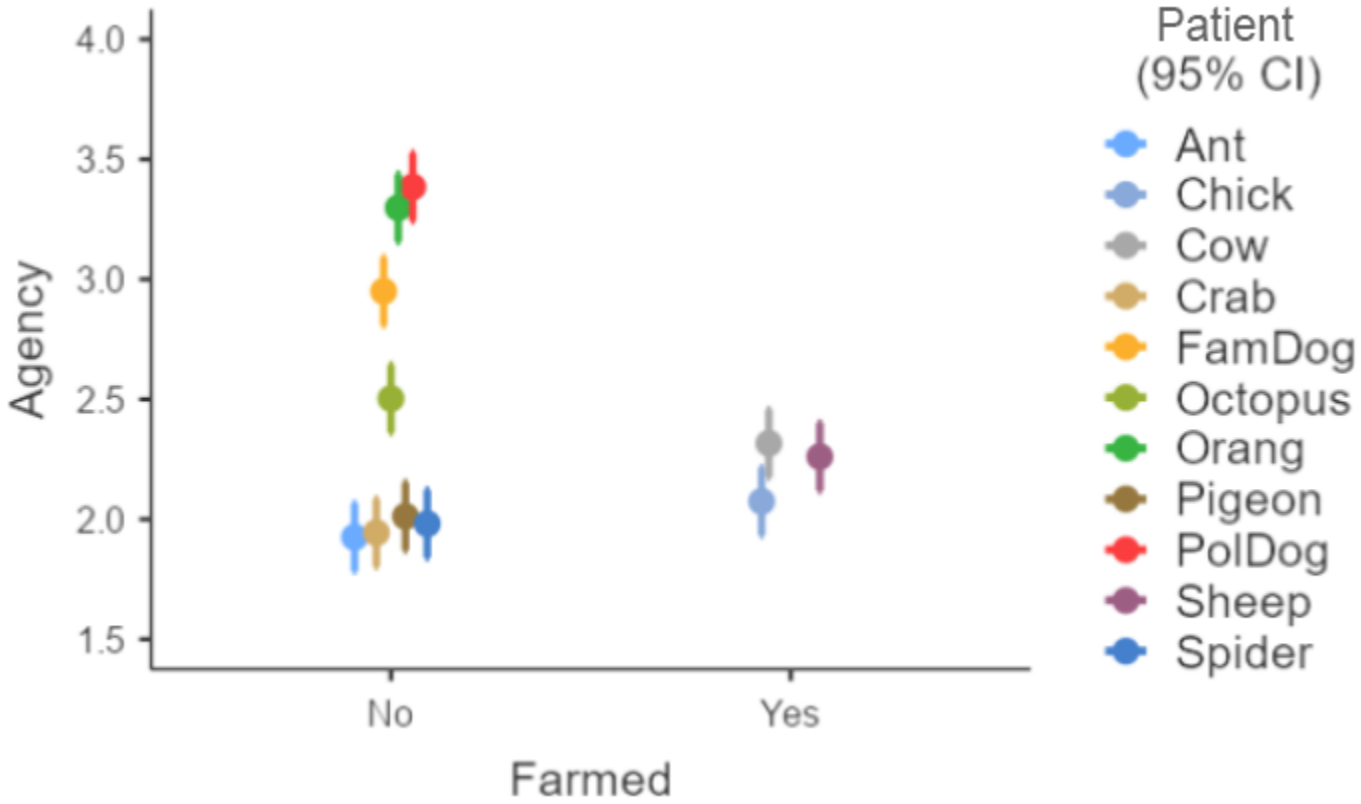


Figure 4.6

Perceived Agency of Farmed and Non-Farmed Animals - Per Animal



Note. *It might be argued that Crab is a farmed animal. This would be consistent as Crab is also rated very low agency.*

Discussion

Once again, people are more likely to think that killing humans is morally worse than torturing them compared to when the same judgments are made for animals. The perceived agency of the target animal/human was found to be the best predictor of these judgments. Some evidence was found for the role of perceived experience but Bayesian analyses indicate that this study was not sensitive enough to draw strong conclusions from this.

Exploratory analyses indicate that farm animals, in particular, are perceived to be very low agency and causing them to suffer was seen as more morally wrong than causing them to die.

It could be argued that one drawback of this study is the within-subjects nature of the design. It is possible that participants make a global judgment about torturing and killing and adjust from there for each patient. This would limit the amount of variance produced from the judgments. The proceeding study will address this.

Experiment 3: Patients Compared Between Subjects

This study uses a between-subjects design with four targets (2 humans and 2 animals) and participants see only one of the targets. The design is otherwise identical to Study 2. This study was pre-registered and can be found here: https://osf.io/bjxpc/

Participants

We recruited 200 participants (aged between 18 and 62, M_{age} = 26.79, SD_{age} = 8.59, 56% male) from Prolific Academic in exchange for a small payment of $2.40.

Procedure

The procedure is the same as Study 2 except for two changes. Participants see only one of four patients (Professor, Person with Disability, Cow, Orangutan). Also, mind perception is measured with twelve questions on a five-point Likert scale (six for agency and six for experience). The agency questions are, *"To what extent do you think the agent: Can communicate with others, is capable of thinking, can plan its actions, is intelligent, has foresight, is able to think things through."* The experience questions are, *"To what extent do you think the agent: Is sensitive to pain, can experience happiness, can experience fear, can experience compassion, can experience empathy, can experience guilt".*

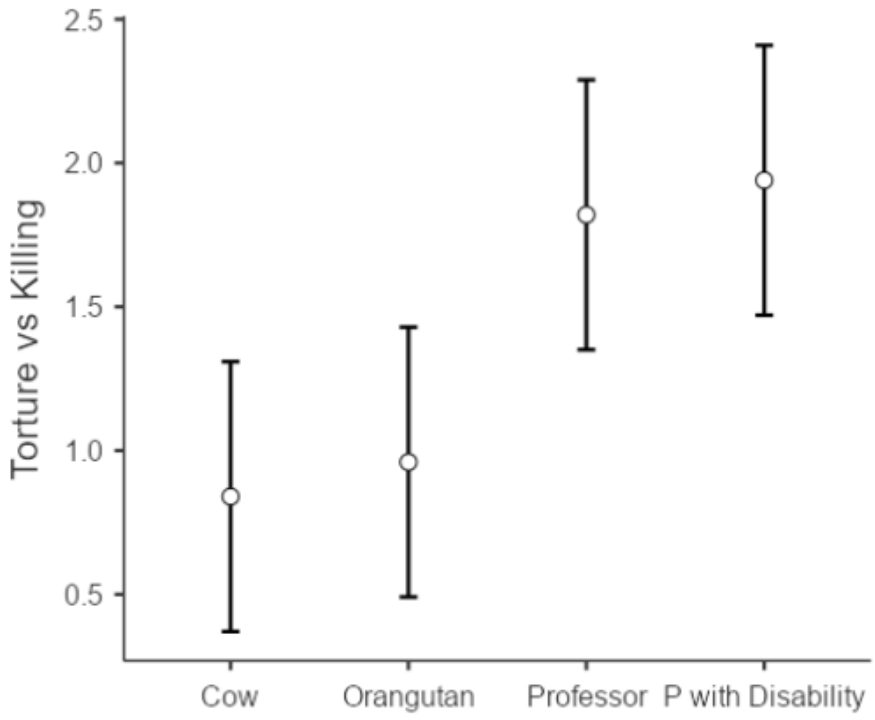
As before, participants are asked which act is more morally wrong on a five-point scale between torture and killing.

Results

To test whether judgments of torture vs killing differed between patients, an independent-groups One-Way ANOVA revealed a significant difference in judgments of torture vs killing (F (3, 108.25) = 5.56, p = 0.001. Games-Howell post hoc comparisons revealed that Animals Cow - Professor p = 0.03; Cow - Person with Disability p = 0.006; Orangutan - Person with Disability p = 0.017, however, there was no significant difference between animals nor between humans. In addition, there was no significant difference between Orangutan - Professor. This can be seen in Figure 4.7.

Figure 4.7

Torture vs Murder for Each Patient



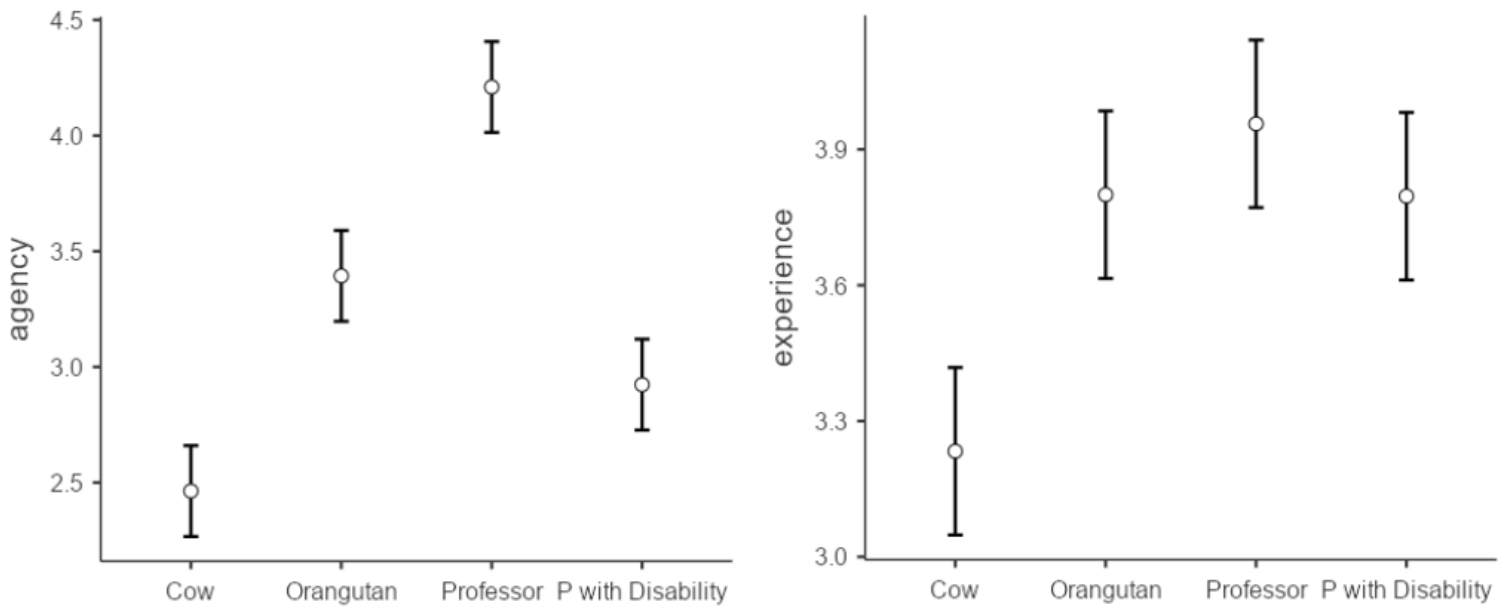
Note. *< 2.5 indicates toture is worse, > 2.5 indicates killing is worse*

To test whether mind perception can account for these judgments a General Linear Model predicting Torture vs Killing with Agency, Experience and Patient-Type (Animal/Human) as predictors. Patient-Type significantly predicted judgments, $F(1,192) = 8.07$, $B = 0.82$, $SE = 0.29$, $p = 0.005$, $BF_{10} = 262.96$;. However, all other predictors were non-significant, Agency, $F(1,192) = 0.002$, $B = -0.01$, $SE = 0.18$, $p = 0.964$, $BF_{10} = 0.20$; Experience, $F(1,192) = 00.02$, $B = -0.03$, $SE = 0.24$, $p = 0.887$, $BF_{10} = 0.17$.

This indicates strong evidence for speciesism and evidence against mind perception. To better visualise how each patient differs in terms of rating of mind perception see Figure 4.8

Figure 4.8

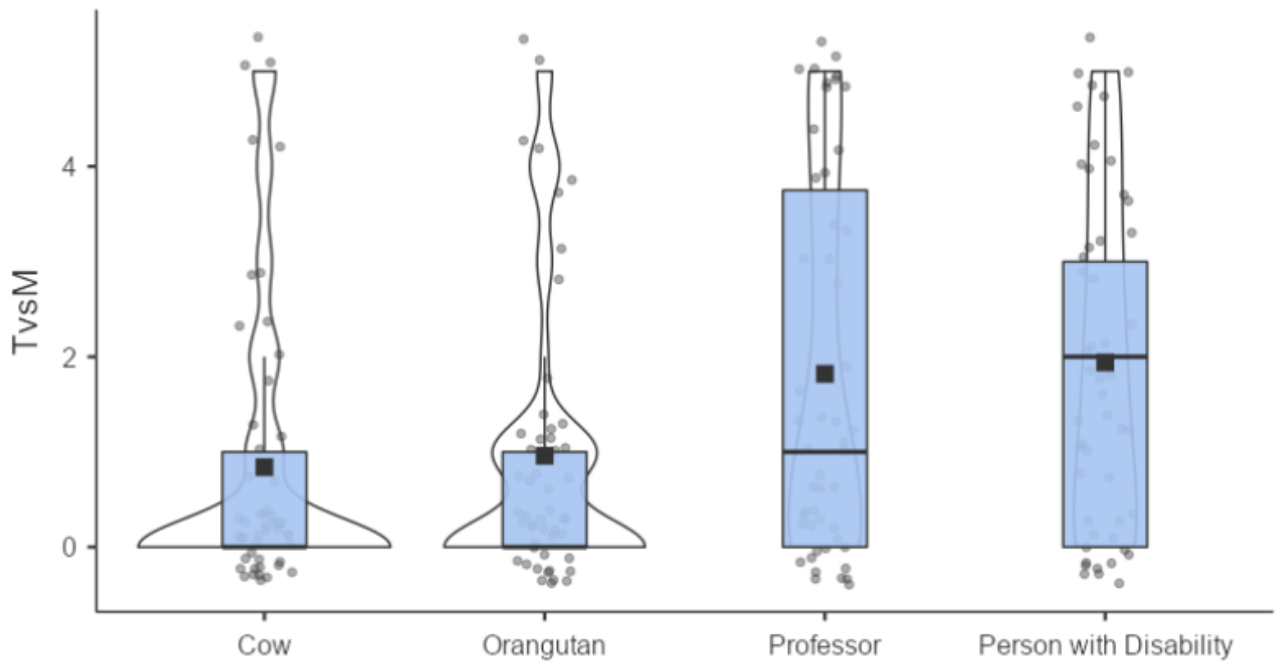
Mind Perception Ratings for each Patient



Once again, torture is rated worse for each patient but in line with the previous studies, the distributions appear to indicate that people were more likely to say that killing is worse than torture for humans compared to animals, see Figure 4.9

Figure 4.9

Torture vs Murder Rating Distributions for each Patient



Note. *< 2.5 indicates torture is worse, > 2.5 indicates killing is worse*

In order to statistically test whether this was the case, responses were re-coded as either "killing is worse" (3 - 5) or "torture is worse (0 - 3). A generalized linear regression was performed to predict recoded torture vs killing judgments with Patient-Type (animal or human) as a predictor. Results confirm that participants were significantly more likely to choose Killing is worse for humans than for animals $\chi^2(1) = 11.88$, B = 1.16, SE = 0.35, $p < 0.001$.

Discussion

Although we found evidence against mind perception accounting for these judgments, it is possible that this may be due to the limited number of targets used in this study compared to the previous study. However, there is clear evidence that speciesism accounts for the data and appears to do so beyond only how we perceive the minds of animals and humans.

The results so far indicate that Speciesism plays a part in how we compare extreme harms, where people are more sensitive to the killing of humans compared to torture but are much less so for animals. The final study explores torture and killing separately to assess how these judgments differ from each other.

So far the results from Study 1 differs from studies 2 and 3 because torturing the specific humans specified in the latter studies were all rated, on average, as worse than killing them. This may be because our mind perception questions do not necessarily measure the ease at which one automatically attributes the kind mind one can empathise with. Rather they may indicate intelectual reflective judgments about the patient. It is possible that the extent we wish to avoid human suffering is driven by the extent to which we perceive and can empathise with a human mind. Specifying a particular human (e.g. a doctor) may increase our ability to empathise with them and therefore wish to avoid their suffering.

Experiment 4: Between Subjects Torture and Killing

So far, animals and humans have consistently generated different ratings when comparing torture to murder in terms of moral wrongness. This experiment separates judgments of torture and killing in order to test the ways in which these judgments differ from each across a selection of pairwise comparisons between targets. In addition to moral wrongness, participants also respond to other questions to get a more rounded view of their moral judgments. This study was pre-registered and can be found here: https://osf.io/ce3kv

Participants

We recruited 160 participants (aged between 18 and 21, M_{age} =18.66, SD_{age} = 0.63, 12% male) from Warwick University's psychology undergraduate student subject pool. Participants received course credit taking part in the experiment.

Procedure

In a between-subjects design, participants are split into either the Killing group or the Toture group. They were presented with the following vignette: *"We need your help! Two criminals are on the loose that have each captured pairs of 7 different animals and humans to [kill/torture]. They intend to [kill their victims painlessly/ torture their victims painfully before letting them go]. Your task is to help us decide which is worse, [killing/torturing] victim A or killing victim B. You will be presented with 21 pairs of these and will answer questions on these pairings. For each one, imagine that this is the ONLY decision you have to make (i.e. no other [killings/torturings] will take place)."*

They are then presented with all possible pairwise comparisons between 8 targets: orangutan, dog, cow, ant, pigeon, octopus, person. For each comparison, participants are asked four questions each on a five-point Likert scale: *"Imagine you have the power to stop ONE of these criminals before they carry this out. Who do you stop?"; "Which act of [killing/torturing] is morally worse?"; "Which criminal is more evil?"; "Imagine they both carry out the killing and are later caught by police. Both receive jail sentences, but one receives a much longer sentence. Who deserves the most punishment?".*
To measure mind perception for each target, eight questions are asked on a five-point scale, four for agency and four for experience. They are asked *"to what extent do you think [patient]: "Is capable of thinking"; "Can plan its actions"; "Is intelligent,"; "Is able to think things through,"; "Is sensitive to pain,"; "Can experience basic emotion (e.g., happiness, fear),"; "Can experience complex emotion (e.g., compassion, guilt)"; "Can experience empathy".*

Results

The pre-registered reliability analysis across the four moral judgments (ie. moral wrongness, stop, evil and punishment) yielded a Cronbach's α = 0.92 indicating that they can be averaged over for the following analyses.

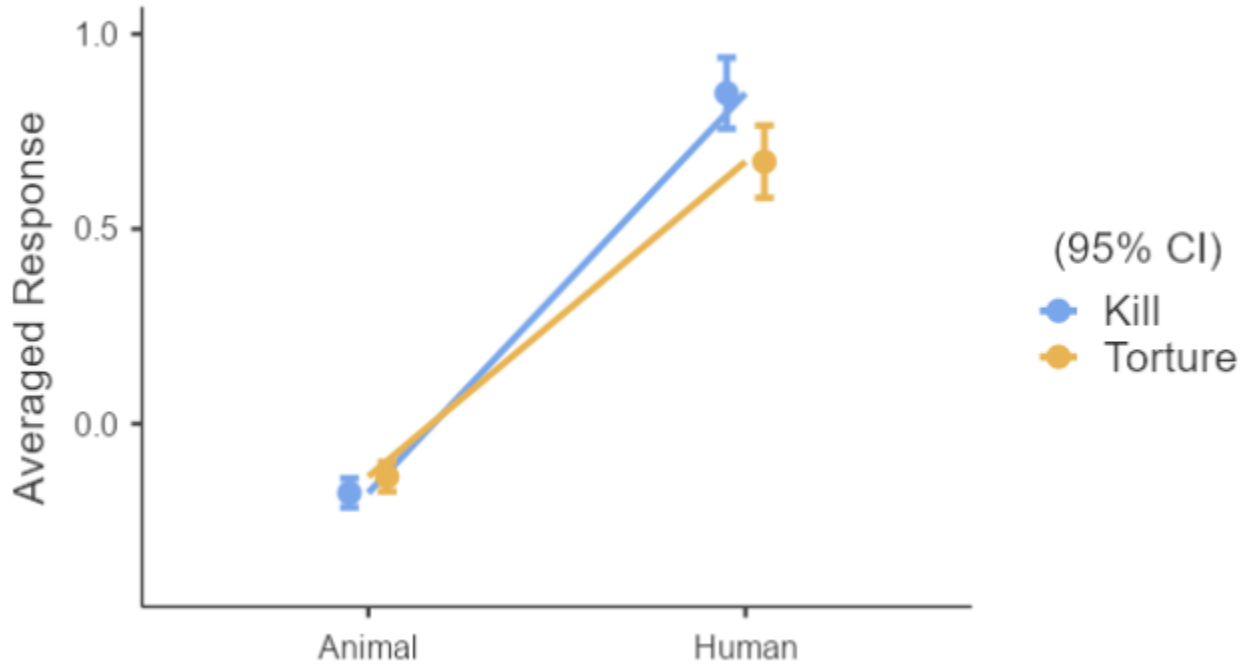
A preregistered linear mixed-model regression was performed to predict average moral judgments with Patient-Type (animal or human), mind perception (perceived Agency and Experience), Group (Torture/Kill) as predictors. (Frequentist: random intercepts and random slopes for Group and Experience by Participant ID).

The results indicate that Patient-Type and Experience were significant predictors as well as the interaction between Group and Patient-Type. Patient-Type: F(1, 7191.33) = 14.51, B = 0.15, SE = 0.04, p < 0.001, BF_{10} = 9.21×10^{10}; Experience: F(1,451.1) = 344.18, B = 0.54 SE = 0.03, p < 0.001, $BF_{10} = 3.08 \times 10^{93}$; Group: F(1, #) = #, B =#, SE = #, p 0.529, BF_{10} = 0.432, p = 0.529; Group by Patient-Type interaction: F(1,3659.13) = 6.09, B = 0.18 SE = 0.07, p = 0.014, BF_{10} = 19.2. Agency, however, was not significant, F(1, 2426) = 0.397, B =-0.1, SE =0.02, p = 0.529, BF_{10} = 0.127.

These are visualised in Figure 4.10. Taken together, these analyses indicate that there is evidence for differences between Patient-Types and evidence that perceived Experience is a predictor of judgments between targets. In addition, there is evidence that judgments between humans and animals differ depending on whether they are made about suffering or killing.

Figure 4.10

Interaction between Torture vs Kill and Patient-Type



Despite Experience and Agency being highly correlated with each other, $p < 0.001$, $r = 0.84$, there is evidence against Agency predicting these judgments. To probe this further, partial correlations were calculated between Agency, Experience and averaged moral judgments. When controlling for Experience, Agency does not significantly correlate with judgments, $p = 0.583$, Pearson's $r = 0.01$. However, when controlling for Agency, Experience does significantly correlate with judgments $p < 0.001$, Pearson's $r = 0.84$.

In light of these results for the following model comparisons, Agency is not included in Mind Perception models.

In order to determine the best model for predicting moral judgments, three Bayesian models were compared. The Speciesism model contained only Patient-Type as a predictor; the Mind Perception model contained only Experience; the full model contained both. Of these models, the best model contained both Experience and Patient-Type, with it performing better than the mind perception model: $BF_{10} = 1.76$; and Speciesism model: $BF_{10} = 1.35 \times 10^{211}$. However, there is only anecdotal evidence that this model is better than the Mind Perception model.

In order to see how the Torture and Killing groups differ from each other, we pre-registered the same analyses for the Killing and Torture groups separately. For each group, linear mixed-model regressions were performed to predict average moral judgments with Patient-Type (animal or human) and mind perception (perceived Agency and Experience) as predictors. (Frequentist: random intercepts and random slopes for Patient-Type and Experience by Participant ID).

The results indicate that for the killing group Patient-Type and Experience were significant predictors while Agency remained non-significant but for the torture group only experience significantly predicted judgments, see table 4.1.

Table 4.1

Comparing predictors for Torture Group and Killing Group

	Killing Group					**Torture Group**				
	B	F	df	p	BF	B	F	df	p	BF
Patient-Type	0.18	5.61	82.46	0.02	3.14×10^{11}	0.093	2.37	3054.6	0.124	6.96
Experience	0.51	0.51	136.32	< 0.001	8.36×10^{59}	0.38	113.33	59.26	< 0.001	3.03×10^{34}
Agency	-0.03	1.13	0.03	0.278	0.556	0.023	0.046	60.1	0.501	0.041

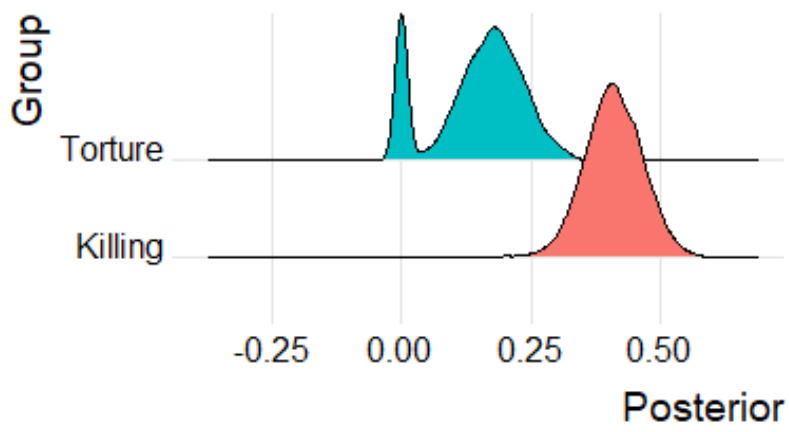
In addition, for the Torture group, adding Patient-Type to the Mind-Perception-Only model failed to substantively improve predictive performance $BF_{10} = 0.043$. In contrast, for the Killing group, adding Patient-Type to the Mind-Perception-Only model decisively increased predictive performance $BF_{10} = 6.64 \times 10^{8}$

Evidence indicates that both Experience and Patient-Type together determine judgments about killing. In other words Speciesism is apparent in killing judgments. However, Mind Perception alone (specifically, judgments about the capacity to feel) accounts for judgments about suffering. This is consistent with the theory that there is Speciesism for judgments about killing but not about suffering.

Finally to assess which theories are most consistent with the data, the weighted posterior 95% HDIs are plotted for the effect of Patient-Type for the Torture and Killing groups, see Figure 4.11

Figure 4.11

Posterior Highest density 95% Credible Intervals for Patient-Type



Taken together, the results of this study are most consistent with the theory that Speciesism strongly affects judgments of killing but are most consistent with no effect, or substantially less effect of Speciesism for judgments about suffering.

Discussion

When making moral judgments about both suffering and killing, the extent to which we think a patient is able to feel and experience internal mental states predicts the strength of these judgments.

Over and above this factor, it also makes a difference whether or not the patient is a human but, importantly, this is specifically the case for judgments about killing. In contrast, when accounting for perceived experience, when making judgments about suffering it does not matter, or matters substantially less, whether the patient is a human or animal. A limitation of this final study is that there is no longer

variation between human patients (for example the lack of inclusion of low-agency humans). This may help explain the differences between the results here and the prior studies.

General Discussion

When comparing how averse we are to either killing or causing suffering to different patients, it appears that there is a complex relationship for the role of Speciesism (the feeling that humans are special) and for the role of mind perception (how we attribute different kinds of minds to different patients). When comparing torturing and killing it appears as if people have asymmetric judgments depending on whether they are considering a hypothetical animal or a human. Participants rated killing as worse for humans and although this was not true when considering individual specific humans or animals, it remained the case that participants were still more likely to rate killing humans as worse than torture than if they were judging animals.

An important factor relating to how we form moral concern for animals is the mental capacities we assign to them. Following the Agency versus Experience model (Gray et al. 2012; Gray et al. 2007; Gray and Wegner 2009) we measured the extent to which participants ascribe mental capacities to the different patients. When people make comparisons between torture and killing, perceived Agency (the ability for an agent to think and act) plays a role in determining which is considered worse but when considering one form of extreme harm in isolation, perceived Experience (the ability for an agent to feel and experience positive and negative mental states) is involved in determining how bad the harm is. It is difficult to understand why this might be the case so future work should explore these capacities in detail. An alternative view of mind perception is provided by Kara Weisman and colleagues where the mental capacities are more represented by "body, heart, and mind" which each relate to aspects of both experience and agency (Weisman et al. 2017; Weisman et al. 2021). This has also been conceptualised as bodily sensation, cognition and, in some cultural settings, emotion (Malle 2021).

It is clear that mind perception plays a role in judgments about killing and suffering and it is also the case that Speciesism exists over and above the kinds of mental abilities we attribute to humans compared to non-human animals (Caviola et al. 2019; Caviola et al. 2020). This is specifically the case when making judgements about killing where we are concerned about the sanctity of human life in general. However, when it comes to mitigating the suffering of different patients what matters most is simply how much we believe those patients are capable of suffering where there is little role for species membership over and above this factor. This was especially true for farm animals who are ascribed lower mental capacities.

We recommend that future work should look at the kinds of arguments animal rights activists appeal to as it is probable that they are more likely to highlight cases of animal suffering rather than appealing to the deaths of animals. In contrast, for policy on human welfare, it is likely that the loss of life will be the most persuasive factor.

Conclusion

Many people are concerned about the suffering that animals endure at our hands, some are even concerned about wild animal suffering. We routinely end animal life, not just for our purposes but also in order to prevent their prolonged suffering. But many people around the world are fighting for the right to end their lives in order to prevent their own prolonged suffering. Animal and human lives are treated categorically differently.

People think humans are special amongst the other species and some have argued this is an entirely

rational position. After all, human minds are particularly complex and we have the ability to be both highly agentic while at the same time having the potential for deep and meaningful phenomenal experience. But even when accounting for these capacities it appears that we have a bias towards human welfare. But what is the shape of his bias? It appears that it predominantly involves the avoidance of the loss of life but much less so the avoidance of suffering. For better or worse, the deontological protections we maintain most strongly for humans centres on life rather than necessarily the quality of that life. When it comes to suffering, we care about the potential for suffering and the extent to which we care about what species the patient largely matches the level of phenomenal experience we ascribe to them.

Chapter 5. Summary and Conclusions

Chapter One introduced the cognitive template, proposed by TDM, for judgments regarding moral violations (Gray and Wegner 2011; Gray et al. 2012; Gray et al. 2014; Schein and Gray 2015; Schein and Gray 2018). To recap, this template contains five elements: A [1. *reaction*] to a *moral* [2. *agent*] who [3. *intentionally*] [4. *violates/harms*] a moral [5. *patient*].

Across three papers, novel empirical work has been conducted to further our understanding of each element. The first paper considers judgments regarding the moral agent and argues that reactions to moral violations can be usefully understood through the lens of Social-Partner Management. By measuring when reactions are either based in Partner Choice or in Partner Control and testing what character-based and act-based judgements relate to these differing strategies, different kinds of violation are clearly associated with a preference for one strategy over the other. The second paper explores how we attribute intentionality to outcomes with and without moral valence, showing that, even though counterfactual reasoning is particularly important, certain models that include counterfactual reasoning fail to accurately predict certain cases. This paper argues that one useful avenue to explore may be causalist models that integrate recent findings from the psychology of causal selection. Finally, the third paper explores the role of mind perception and speciesism when making judgments regarding the moral patient. Building on previous work showing that mind perception alone cannot explain moral-worth judgments, confirming speciesist prejudices in judgments, the work conducted here illustrates this is fundamental for judgments of killing but much less so for judgments of suffering.

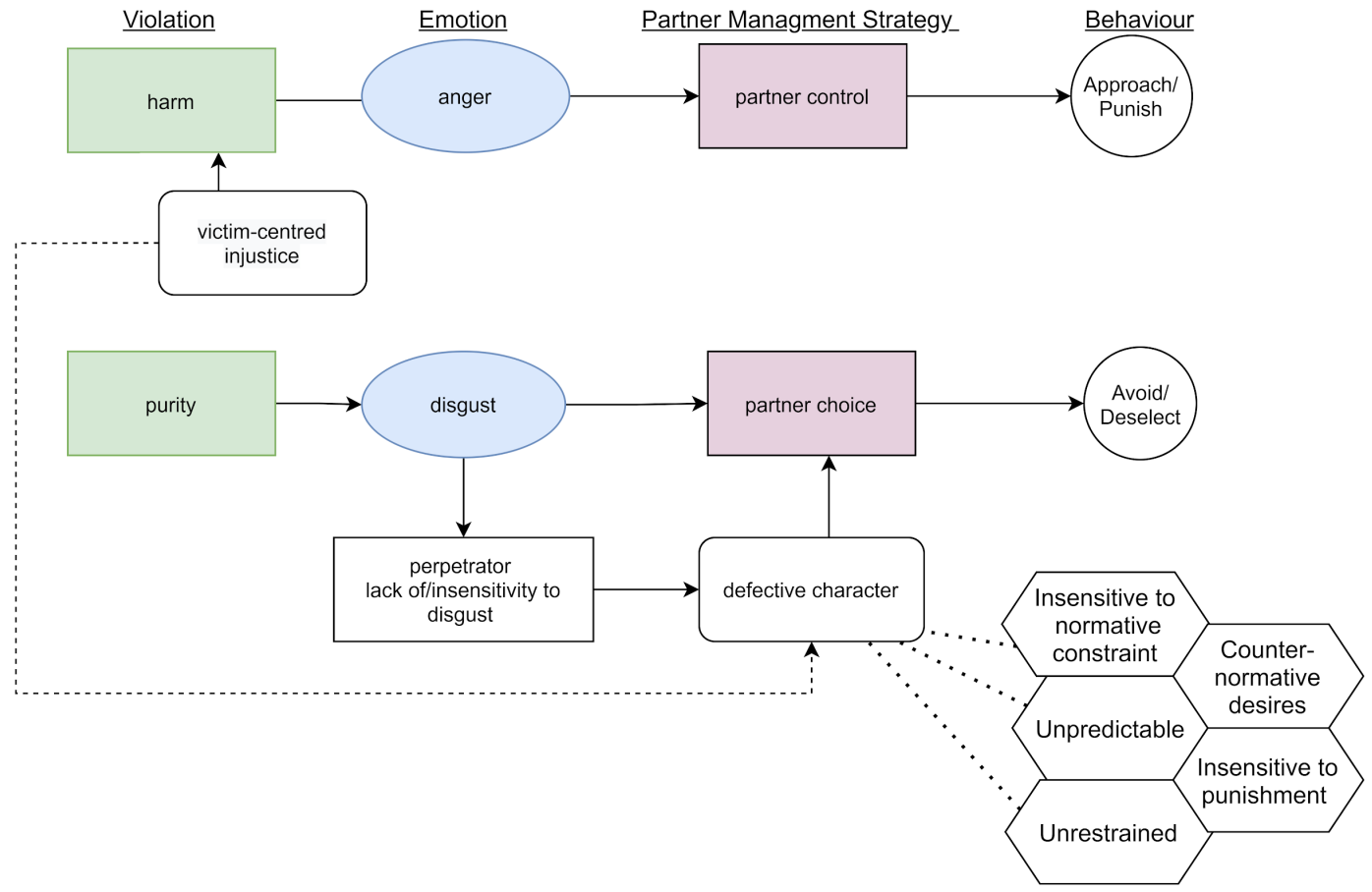
Work trying to understand how we react to moral transgressions dominate the field of moral psychology (issues concerning 1 and 4). We react in different ways to the different kinds of moral transgression that make up the purportedly separate and distinguishable categories that partition the moral domain (Haidt and Joseph 2008; Piazza et al. 2018; Graham et al. 2018; Curry et al. 2019; Curry 2016). Discovering and operationalizing these categories has allowed us to better understand individual differences in our moral attitudes, especially across the political divide (Janoff-Bulman and Carnes 2013; Waytz et al. 2019; Kivikangas et al. 2020; Frimer et al. 2015; Smith et al. 2019; Graham et al. 2009). It has been demonstrated on multiple occasions that Liberals and Conservatives appear to care about different categories of morality and reactions to these categories are associated with different affective systems, and this finding was replicated in the empirical work conducted in Chapter 2. The pressing question remains as to why this is the case. To shed light on this, Chapter 2 argues that the way we react to different moral transgressions might be best understood through the lens of Social-Partner Management. This term is adopted here to cover the contrasting behavioural strategies of Partner Choice and Partner Control (Bull and Rice 1991; Campennì and Schino 2014; Noë 2006; Martin et al. 2020; Barclay and Raihani 2016; Martin et al. 2019). For moral transgressions, Partner Control relates to interpersonal, 3rd-party or institutional punishments which are intended to prevent the undesirable behaviours of social partners. Partner Choice, for moral transgressions, relates to avoiding and/or deselecting those who have shown less willingness to cooperate and appear to be less beneficial partners. Importantly, Partner Choice is argued to rely on Character-Based Judgments (concerning 2 in the cognitive template). Character-Based Judgments (e.g. Uhlmann et al. 2013; Uhlmann et al. 2015; Pizarro et al. 2003; Everett et al. 2016; Pizarro and Tannenbaum 2012; 2011) relate to the signals and inferences over the moral character of the agent or transgressor. Here, this especially relates to the utility those inferences provide for social partnerships and group cohesion.

Chapter 2 explored how different kinds of moral transgression lead to a preference for either Partner Choice or Partner Control. We specifically focused on violations of harm and purity from MFT because they are the most widely contrasted and also most associated with individual differences in sensitivity (Graham et al. 2009; Kivikangas et al. 2020). Across two experiments, harm violations reliably lead to more Partner Control reactions and purity violations reliably lead to more Partner Choice reactions. This was true even when considering third-party punishment. The results indicate that when people are presented with situations where a victim is clearly harmed, they make judgments of moral wrongness, they are most likely to feel anger (an approach-based emotion), and are most likely to seek outcomes related to punishment. In contrast, when people are presented with purity violations, where a particular moral patient is not easily identifiable, people are more likely to feel disgust (an avoid-based emotion), and seek outcomes related to the avoidance of that perpetrator. When reacting with Partner Control, the strength of this reaction was most associated with how morally wrong the action was judged to be, while Partner Choice reactions were more associated with how uncomfortable people would feel around the violator. This was true even if it was understood that the perpetrator had been sufficiently punished. In addition, purity violators were seen as less predictable and participants found their motivations more difficult to understand. When assessing emotional reactions, harm violations were more associated with anger (and guilt when one imagines themselves as the perpetrator) while purity violations were more associated with disgust (and shame if they were the perpetrator). All these judgments appear to moderate the strength of the Partner Management strategy but they do not themselves appear to determine which strategy is preferable.

Based on these results, Figure 5.1 outlines a proposed model of how judgments of purity and harm violations may lead to different social partner management strategies. The green boxes note either a harm or a purity violation. Harms result in anger, leading to Partner Control, specifically punishment, either interpersonal or institutional. In contrast, observing a purity violation leads to disgust, potentially driven by inferences over the character of the perpetrator (for example, driven by observing their lack of disgust at performing an otherwise disgusting actor or their desire to participate in counternormative activities). Some options are offered by the interlocking diagrams with regard to the kinds of character-based inferences that may be made, for example, the agent may be perceived to be unpredictable and therefore an undesirable social partner. Alternatively, Partner Choice (avoidance of these social partners) may be more affectively driven rather than requiring these kinds of explicit judgments (Haidt 2001)

Figure 5.1.

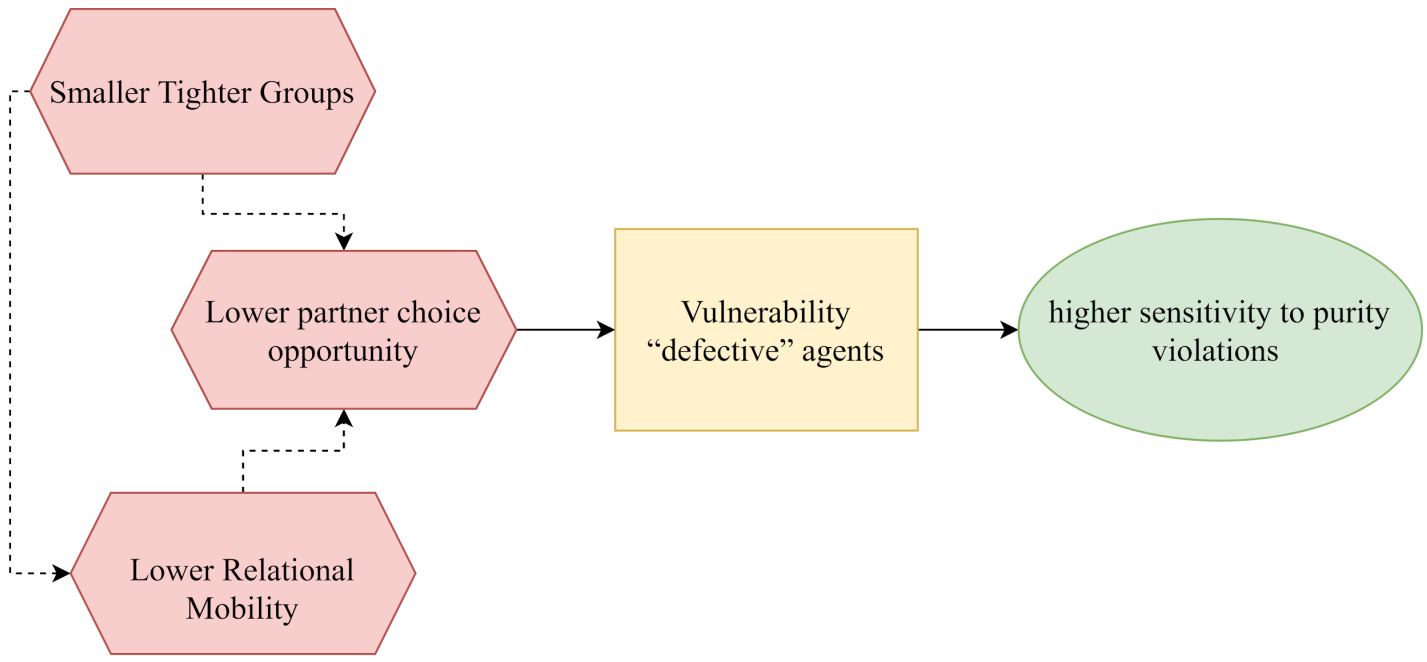
Partner Management Process Diagram



The second part of the empirical work conducted in Chapter 2 relates to individual differences in these strategies. Unexpectedly, no individual differences measures were associated with differences in Partner Management reactions. This was surprising considering that liberals and conservatives differ in their sensitivity to purity and harm (Kivikangas et al. 2020; Graham et al. 2009), a finding which was also replicated in our sample (Experiment 1). A potential model for individual differences is illustrated in figure 5.2. Differences across the political divide are taken to be driven by differences across group structures. Conservatives are more likely to be part of smaller and tighter communities and social groups (Maxwell, 2019; Waytz et al., 2019), this being especially the case across the rural vs urban divide (Nemerever and Rogers 2021), where people have less Relational Mobility (Awad et al. 2020; Thomson et al. 2018). In such groups, one may be more vulnerable to potentially harmful agents (since they can't be avoided and are more likely to be social partners in the social network), who although may have not committed a harm yet may be believed to have the potential to cause harm in the future (Chakroff 2015). Therefore, in such groups, it may be more beneficial to be sensitive to strong character-based signals such as those inferred from purity violations. The lack of individual differences in Partner Management in the results of Chapter 2 is difficult to take at face value as the three constructs (Political Orientation, Urban vs Rural living situation, Relational Mobility) were not associated in any way. Despite this, there remains good reason to believe that such an association exists (e.g. Nemerever and Rogers 2021) which may indicate an issue with the sample or an issue with the measures used. Therefore future research should explore this further.

Figure 5.2.

Sensitivity to Purity Violations



Since the SEE was discovered almost twenty years ago, issues pertaining to intentionality ascription (the third element in the cognitive template) have been widely debated across both philosophy and psychology journals. With little chance to resolve the debate as a whole, the second paper addresses one clearly important aspect, that of counterfactual reasoning, (Halpern and Kleiman-Weiner 2018; Kleiman-Weiner et al. 2015; Quillien and German 2021). A particular previously successful account argues that secondary outcomes are judged to be intentional if they are counterfactually required for a desired outcome to be achieved (Halpern and Kleiman-Weiner 2018; Kleiman-Weiner et al. 2015). This model has been useful as it can predict judgments to particular kinds of trolley problems and has also been integrated into models of blame ascription. Usefully, it also makes novel predictions about what people may infer from classic SEE cases. If judgements involving counterfactual requirement intervene on the classic SEE case then people may make different assumptions about harm cases compared to help cases (that harm cases are counterfactually required but help cases are not). However, the results demonstrated that this was not the case (Experiment 1). Although this means that the simple counterfactual requirement model fails to explain the SEE, it is important to note that it was not intended to do so. However, it is possible to modify the vignettes such that, if counterfactual requirement was integral to intentionality ascription, it should intervene on these modified cases. Despite this, when it is made explicit whether the secondary outcomes are counterfactually required for a policy to make money, we observe no significant difference in ascriptions of intentionality(Experiment 2).

We explored what this means for the IDH which states that findings in this field are best explained by there being multiple different concepts activated by the word "intentional" (Laurent et al. 2020; Laurent et al. 2019; Nichols and Ulatowski 2007; Cova ; Cushman and Mele 2008). However, it is not clear that any IDH account currently explains these findings. Counterfactual reasoning almost certainly plays a role in these judgements so what might explain this discrepancy? The paper draws attention to very recent work linking intentionality judgments to our most up to date models of causal reasoning (Quillien and German 2021; Quillien 2020). Rather than centering on how an individual counterfactually maps reality, these models focus on how these maps (infrared foreknowledge, desires, expected utility calculations) are determined to be causative of the outcomes. Causal selection is itself driven by counterfactual reasoning and thus ought to be explored further. For example, future work may shed light on why the Counterfactual Requirement model predicts trolley cases but not the classic or modified-classic SEE cases.

Finally, the third paper focuses on how judgments are affected by considerations regarding the moral patient (the fifth element of the cognitive template). Previous work has shown that the way we perceive the minds of a moral patient affects our moral judgement (Schein and Gray 2018; Gray et al. 2012; Caviola et al. 2020; Caviola et al. 2019). In particular, we appear to hold humans in especially high regard compared to our animal cousins. The dimensions of mind perceptions are partly able to explain this but being a member of the human species seems to bestow extra moral protection over and above this (Caviola et al. 2020; Caviola et al. 2019). However, the question arises as to whether this is the case for all kinds of moral violation. Chapter 4 explores this across four experiments. Participants were asked compare acts of torture and acts of killing but importantly some were considering animal victims where others were considering human victims. It appears as if people are more likely to rate killing as being morally worse than torture for humans, but they are less likely to do so for animal victims. Mind perception with regards to the target was able to partly explain these findings but over and above this Speciesism also appears to matter for comparing these different kinds of harm.

However, the role of Speciesism is complex. When making judgements about killing, we are particularly concerned about the sanctity and protection of human life in general but when it comes to mitigating suffering, mind perception is dominant. In other words, for judgments over torture, what matters the most is how much we believe those patients are capable of suffering (perceived patiency/experience). Being human is less likely to confer any extra privilege in these cases. The deontological constraints we maintain appear to protect humans from death, but there is less evidence that deontological constraints for humans compared to animals, protect the quality of that life. When it comes to suffering, we care about the potential for suffering, whatever the species, and we infer the potential for suffering based on the phenomenal experience we ascribe to the moral patient. Based on these findings, it is likely that animal rights activists would appeal to the prevention of animal suffering rather than appealing to the deaths of animals but for policy on human welfare, it is likely that people we are most persuaded by highlighting the potential loss of life.

Using the cognitive template suggested by TDM, this thesis has partitioned judgments over moral violations into separate elements that make up a paper delving into the complexity of reactions to transgressions, a paper covering speciesism and a paper covering intentionality ascription. Intentionality plays a crucial rose in both moral judgement and social cognition more generally. Moreover, counterfactual reasoning is fundamental to this and also to many other areas of cognition. A full-fledged account of counterfactual reasoning may help solve issues with intentionality ascription that have plagued the field for almost twenty years and a successful integrated model of intentionality ascription may also hold the key to understanding moral judgement on a more fundamental level. In terms of reactions to moral transgressions, there is still debate with regards to what constitutes a moral transgression and how simple harms differ from more heterogeneous purity violations. The paper offers an important step in understanding moral judgements as a form of social partner management. In addition, understanding how and why people react in the way they do can help inform researchers and policymakers interested in how we structure our social hierarchies and institutions and can help us understand how individuals can coexist and cooperate in the social world. It can also shed light on the potential reasons why there is so much disagreement on this across political lines. Finally, many people are concerned about the suffering of non-human animals. The speciesism paper explores fundamental differences between the ways we think about ourselves and our cousins in the animal kingdom. Although people may think humans are special amongst the other species, whether this is a bias and what shape this bias takes can also provide insight into why there is resistance to allowing individuals the right to end their own lives under certain

circumstances (for example to end prolonged suffering). It is our hope that these disparate avenues of research can be integrated into a more detailed and rich picture of how we react to wrong-doers

Glossary of Terms

Act Utilitarianism

A normative theory that states that the moral rightness of an act reduces to the consequences which that act brings about in that situation. Acts that bring about the maximum utility are the acts one most ought to do. Utility can be defined in many ways, see for example **Hedonistic Utilitarianism** or **Desire Utilitarianism** in *Glossary*. For further details see: (Driver 2014).

Act-Based Judgement

In the psychology of moral judgement, act-based judgments focus purely on the act itself, whether it is deontologically prohibited, or prohibited based on the negative utility it brings about or some other concern about the act itself. This is contrasted with **Character-Based** judgments that focus instead on the inferred moral-character of the agent.

Agency

One of the two dimensions of **Mind Perception**. When attributing mind to different entities, the kind of mind one perceives will vary across two theoretical constructs. Agency is the entity's ability to think, plan, use language and act on the world. The dimensions of mind perception also reflect the two archetypes of a moral dyad. The agent (in cases of moral violation this would be the perpetrator) and the patient (in cases of moral violation this would be the victim). See also **Experience** and **Mind Perception** in *glossary*.

Anthropocentrism

The belief or bias that places humans at the centre of all concerns. For example, inferring that natural phenomena are designed for human purposes, or the view that human beings are supreme, or the tendency to view other species by analogy to humans.

Approach/Avoid

Behavioural regulation for appetitive and aversive stimuli, respectively. The Behavioural Activation System (BAS) activativates approach-based behaviours where an agent would move towards the target. The Behavioural Inhibition System (BIS), in contrast, inhibits behaviours that would do so, in order to avoid the target.

Causelist Theory

In Philosophy of Action, causelist theories of intentionality state that an outcome is intentional in so far as the acting agent's propositional attitudes (desiring, knowing, fearing, believing etc) cause the outcome in question.

Character-Based Judgement

Also known as "Personological" judgement. In moral psychology, moral judgments made on the basis of how much one can infer about the moral character of the agent are character-based judgments. These are contrasted with **Act-Based Judgments** that only consider the act and its consequences.

Deontological

The normative theory which states that the rightness and wrongness of an act is determined by a set of rules and not only the consequences of the act. The nature and origins of the rules differ between Deontological theories, e.g. acts can intrinsically be good or bad in-themselves, or they exist by divine command, or they result from a set of prerequisite duties etc.

Experience

Also known as **Patiency**. One of the two dimensions of **Mind Perception**. When attributing mind to different entities, the kind of mind one perceives will vary across two theoretical constructs. Experience is the entity's ability to have internal phenomenal experiences like pain and pleasure or fear and hunger etc. The dimensions of mind perception also reflect the two archetypes of a moral dyad. The agent (in cases of moral violation this would be the perpetrator) and the patient (in cases of moral violation this would be the victim). See also **Agency** and **Mind Perception** in *glossary*.

Hedonistic Utilitarianism

A moral theory that argues that the principle of utility is grounded in the consequential mental states produced by an act, (see **Act Utilitarianism)**. This means that the unit of utility one seeks maximise, or the unit by which acts are measured is happiness, satisfaction and/or the reduction of pain or displeasure.

IDH - Interpretive Diversity Hypothesis

Due to the difficulty of finding a single model that explains all the data on folk intentionality, this view posits that the word “intentional” actually captures separate and distinct concepts and one infers what concept is being asked about from the context of the question. For example, it may be asking about foreknowledge on the one hand or intent on the other.

Knobe Effect

Named after the Philosopher Joshua Knobe, who discovered the effect, this term relates to the highly replicated finding that people demonstrate strange asymmetries in their ascriptions of intentionality. Specifically, people seem to make these judgments using evaluative or normative reasoning rather than simply reasoning about the mental states of the agent. See also **SEE.**

Mind Perception

People automatically ascribe minds to different entities but the nature of the mind one perceives may differ across different dimensions. TDM argues that for moral judgement to take place, two causally connected minds must be perceived. See also **Agency**, **Experience, TDM**.

Moral Agent

In a moral dyad, the moral agent is the acting agent, the agent which affects the world, or specifically affects the moral patient. In moral transgressions the moral agent harms or violates the patient in some way. See also **TDM.**

Moral Patient

In a moral dyad, the moral patient is the one who is acted upon, the one who is affected. For moral transgressions the moral patient is the one who is harmed or violated. See also **TDM**.

MVF - Moral Foundations Theory

This is the pluralistic view that morality consists of separate categories or foundations. The original formulation of MFV identifies five such categories, Harm, Fairness, Authority, Loyalty and Purity. However, this is not argued to be an exhaustive list and the theory permits other possibilities (for example Liberty).

Partner Choice

Behaviours intended to allow the formation of social partnerships, joint action and cooperation, or dissolve current social partnerships and reject potentially undesirable partners. Individuals

attempt to outbid each other in the biological market in order to benefit from joint cooperation. This is contrasted with **Partner Control**.

Partner Control

Behaviours intended to reinforce desirable behaviours or prevent undesirable ones. For example, through directly rewarding or punishing individuals. This is contrasted with **Partner Choice**.

Patiency

See **Experience**

Personological

See **Character-Based Judgement.**

Preference Utilitarianism

A moral theory that argues that the principle of utility is grounded in whether an individual's preferences are fulfilled or frustrated by the act (see **Act Utilitarianism)**. This means that the unit of utility one seeks maximises, or the unit by which acts are measured is the extent to which it aligns with the agent's preferences.

Purity

One of the five identified categories in **Moral Foundations Theory**. This category is the most widely debated. Purity violations can range from acts related to sex, disordlyness, contamination or spiritual deflilment

Relational Mobility

A theoretical construct intended to measure how much freedom and opportunity a society affords people to form new social relationships and dissolve current ones.

SE - Skill Effect

In intentionality ascription, this effect happens when people compare skillful actions to unskilful actions (where desired outcomes occur mostly through luck). When making judgments that are morally salient people appear to disregard skill in favour of evaluations based on the desires of the agent.

SEE - Side Effect Effect

See **Knobe Effect**

Speciesism

The belief or bias that Humans are in some-way special and of greater importance or concern than our animal cousins.

TDM - Theory Of Dyadic Morality

The psychological theory that moral judgement is framed around a cognitive template of a moral agent harming a moral patient. This is a monist view that all moral judgement revolves around harm and requires the **mind perception** of two causally connected agents, one acted upon (the **moral patient**) and one performing the action (the **moral agent**).

References

Alicke, M. (2008). Blaming Badly. *Journal of Cognition and Culture*, *8*(1-2), 179–186.

Allen, C., & Trestman, M. (2017). Animal Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/consciousness-animal/

Amin, A. B., Bednarczyk, R. A., Ray, C. E., Melchiori, K. J., Graham, J., Huntsinger, J. R., & Omer, S. B. (2017). Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, *1*(12), 873–880.

Appiah, K. A. A. (2007). *Ethics in a World of Strangers*. https://edoc.hu-berlin.de/bitstream/handle/18452/2372/vorlesung_appiah_kwame_ethics.pdf?sequence=1

Arico, A. J. (2012). Breaking Out of Moral Typecasting. Review of Philosophy and Psychology, 3(3), 425–438.

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(5), 2332–2337.

Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology*, *7*, 33–38.

Barclay, P., & Raihani, N. (2016). Partner choice versus punishment in human Prisoner's Dilemmas. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *37*(4), 263–271.

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). https://doi.org/10.1016/j.jml.2012.11.001

Batson, C. D., Ahmad, N., Lishner, D. A., Tsang, J., Snyder, C. R., & Lopez, S. J. (2002). Empathy and altruism. *The Oxford Handbook of Hypo-Egoic Phenomena*, 161–174.

Bentham, J. (1789). A utilitarian view. *Animal Rights and Human Obligations*, 25–26.

Bernhard, R. M., LeBaron, H., & Phillips, J. S. (2021). *It's not what you did, it's what you could have done*. https://doi.org/10.31234/osf.io/gb3q7

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding Robots Responsible: The Elements of Machine Morality. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.02.008

Björklund, F., Haidt, J., & Murphy, S. (2000). *Moral dumbfounding: when intuition finds no reason*. https://portal.research.lu.se/portal/en/publications/moral-dumbfounding-when-intuition-finds-no-reason(a2e90d4a-3655-433f-a186-fd6f8ca5bccd).html

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of Mice, Men, and Trolleys: Hypothetical Judgment

Versus Real-Life Behavior in Trolley-Style Moral Dilemmas. *Psychological Science, 29*(7), 1084–1093.

Bull, J. J., & Rice, W. R. (1991). Distinguishing mechanisms for the evolution of co-operation. In *Journal of Theoretical Biology* (Vol. 149, Issue 1, pp. 63–74). https://doi.org/10.1016/s0022-5193(05)80072-4

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Burra, A., & Knobe, J. (2006). The Folk Concepts of Intention and Intentional Action: A Cross-Cultural Study. In *Journal of Cognition and Culture* (Vol. 6, Issues 1-2, pp. 113–132). https://doi.org/10.1163/156853706776931222

Byrne, R. M. J. (2017). Counterfactual Thinking: From Logic to Morality. *Current Directions in Psychological Science*, *26*(4), 314–322.

Campennì, M., & Schino, G. (2014). Partner choice promotes cooperation: the two faces of testing with agent-based models. *Journal of Theoretical Biology, 344*, 49–55.

Carson, T. L. (1983). Utilitarianism and the Wrongness of Killing. *Erkenntnis. An International Journal of Analytic Philosophy*, *20*(1), 49–60.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: evidence and implications. *Psychological Bulletin*, *135*(2), 183–204.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of Personality and Social Psychology, 67*(2), 319.

Caviola, L., Everett, J. A. C., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology, 116*(6), 1011–1029.

Caviola, L., Kahane, G., Everett, J. A. C., Teperman, E., Savulescu, J., & Faber, N. S. (2020). *Utilitarianism for animals, Kantianism for people? Harming animals and humans for the greater good*. https://doi.org/10.1037/xge0000988

Chakroff, A. (2015). *Discovering Structure in the Moral Domain*. https://dash.harvard.edu/handle/1/17467227

Chakroff, A., Dungan, J., & Young, L. (2013). Correction: Harming Ourselves and Defiling Others: What Determines a Moral Domain? *PloS One, 8*(9). https://doi.org/10.1371/annotation/38818ce6-1b40-4965-aa64-b7943d2711ed

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: an attribution asymmetry across moral domains. *Cognition*, *136*, 30–37.

Chapman, H. A., & Anderson, A. K. (2013). Things rank and gross in nature: a review and synthesis of moral disgust. *Psychological Bulletin*, *139*(2), 300–327.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: a standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research*

Methods, *47*(4), 1178–1198.

Coke, J. S., Batson, C. D., & McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology*, *36*(7), 752–766.

Cooley, E., Payne, B. K., Cipolli, W., III, Cameron, C. D., Berger, A., & Gray, K. (2017). The paradox of group mind:"People in a group" have more mind than "a group of people." *Journal of Experimental Psychology. General*, *146*(5), 691.

Corr, P. J. (2002). J. A. Gray's reinforcement sensitivity theory: tests of the joint subsystems hypothesis of anxiety and impulsivity. *Personality and Individual Differences*, *33*(4), 511–532.

Cova, F. (2016). The Folk Concept of Intentional Action: Empirical Approaches. In W. Buckwalter & J. Sytsma (Eds.), *Blackwell Companion to Experimental Philosophy*.

Cova, F., & Naar, H. (2012). Side-Effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality. *Philosophical Psychology*, *25*(6), 837–854.

Curry, O. S. (2016). Morality as Cooperation: A Problem-Centred Approach. In T. K. Shackelford & R. D. Hansen (Eds.), *The Evolution of Morality* (pp. 27–51). Springer International Publishing.

Curry, O. S., Jones Chesters, M., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of "morality-as-cooperation" with a new questionnaire. *Journal of Research in Personality*, *78*, 106–124.

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., & Mele, A. (2008). Intentional action: Two-and-a-half folk concepts? In J. Knobe (Ed.), *Experimental philosophy , (pp* (Vol. 244, pp. 171–188). Oxford University Press, xi.

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford Handbook of Moral Psychology*, 47–71.

Dablander, F. (2020). *An Introduction to Causal Inference*. https://doi.org/10.31234/osf.io/b3fkw

Dalbauer, N., & Hergovich, A. (2013). Is What is Worse More Likely?—The Probabilistic Explanation of the Epistemic Side-Effect Effect. *Review of Philosophy and Psychology*, *4*(4), 639–657.

Damon, W. (1999). The moral development of children. *Scientific American*, *281*(2), 72–78.

Darnton, R. (2009). *The great cat massacre: And other episodes in French cultural history*. Basic Books.

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy*, *60*(23), 685–700.

Davidson, D. (1980). *Essays on Actions and Events*. Clarendon Press.

Derbyshire, E. J. (2016). Flexitarian Diets and Health: A Review of the Evidence-Based Literature. *Frontiers in Nutrition*, *3*, 55.

Descartes, R. (1989). The passions of the soul (1649). *The Philosophical Writings of Descartes*, *1*, 11.

Ditto, P. H., & Jemmott, J. B., 3rd. (1989). From rarity to evaluative extremity: effects of prevalence

information on evaluations of positive and negative characteristics. *Journal of Personality and Social Psychology*, *57*(1), 16–26.

Driver, J. (2014). The History of Utilitarianism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2014). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history/

Dungan, J. A., Chakroff, A., & Young, L. (2017). The relevance of moral norms in distinct relational contexts: Purity versus harm norms regulate self-directed actions. *PloS One*, *12*(3), e0173405.

Epley, N., & Waytz, A. G. (2009). Perspective taking. In *Encyclopedia of human relationships* (pp. 1228–1231). Sage Publications, Inc.

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology. General*, *145*(6), 772–787.

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science*, *24*(1), 56–62.

Feinberg, M., & Willer, R. (2019). Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, *13*(12), 29.

Fetherstonhaugh, D., Slovic, P., Johnson, S., & Friedrich, J. (1997). Insensitivity to the Value of Human Life: A Study of Psychophysical Numbing. *Journal of Risk and Uncertainty*, *14*(3), 283–300.

Fincher, C. L., & Thornhill, R. (2012). The parasite-stress theory may be a general theory of culture and sociality [Review of *The parasite-stress theory may be a general theory of culture and sociality*]. *The Behavioral and Brain Sciences*, *35*(2), 99–119.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906.

Fjellstrom, R. (2002). Specifying Speciesism. *Environmental Values*, *11*(1), 63–74.

Frimer, J. A., Tell, C. E., & Haidt, J. (2015). Liberals condemn sacrilege too: The harmless desecration of Cerro Torre. *Social Psychological and Personality Science*. http://journals.sagepub.com/doi/abs/10.1177/1948550615597974

Gilligan, C. (1994). In a different voice: Women's conceptions of self and of morality. *Caring Voices and Women's Moral Frames: Gilligan's View. Moral Development: A Compendium*, *6*, 1–37.

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond Purity. *Psychological Science*, *28*(1), 80–91.

Goodman, J. R., Borch, C. A., & Cherry, E. (2012). Mounting Opposition to Vivisection. *Contexts* , *11*(2), 68–69.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Academic Press.

Graham, J., Haidt, J., Motyl, M., Meindl, P., Iskiwitch, C., & Mooijman, M. (2018). Moral foundations theory. *Atlas of Moral Psychology*, 211–222.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385.

Grau, C. (2016). A Sensible Speciesism? *Philosophical Inquiries*, *4*(1), 49–70.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Gray, K., DiMaggio, N., Schein, C., & Kachanoff, F. (2021). *What is "purity"? Conceptual murkiness in moral psychology*. https://doi.org/10.31234/osf.io/vfyut

Gray, K., & Keeney, J. E. (2015). Impure or Just Weird? Scenario Sampling Bias Raises Questions About the Foundation of Morality. *Social Psychological and Personality Science*, *6*(8), 859–868.

Gray, K., MacCormack, J. K., Henry, T., Banks, E., Schein, C., Armstrong-Carter, E., Abrams, S., & Muscatell, K. A. (2022). The affective harm account (AHA) of moral judgment: Reconciling cognition and affect, dyadic morality and disgust, harm and purity. Journal of Personality and Social Psychology. https://doi.org/10.1037/pspa0000310

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology. General*, *143*(4), 1600–1615.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*(3), 505–520.

Gray, K., & Wegner, D. M. (2010). Blaming god for our pain: human suffering and the divine mind. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *14*(1), 7–16.

Gray, K., & Wegner, D. M. (2011). Morality takes two: Dyadic morality and mind perception. *The Social Psychology of Morality: Exploring the Causes of Good and Evil*, *2011*, 109–127.

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124.

Greene, J. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. (2016). Solving the trolley problem. *A Companion to Experimental Philosophy*. https://books.google.com/books?hl=en&lr=&id=kezbCwAAQBAJ&oi=fnd&pg=PA175&dq=Solving+Trolley+Problem+Greene&ots=G4cJIe7utX&sig=6hSyMazAXpNwUpBPBM7yQ2t5new

Greene, J. D. (1294). The cognitive neuroscience of moral judgment. *The Cognitive Neurosciences., 4th Ed.*, *4*(2009), 987–999.

Gruen, L. (2017). The Moral Status of Animals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*

(Fall 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/moral-animal/

Gruen, L., & Jones, R. C. (2015). *Veganism as an Aspiration*. https://www.wellbeingintlstudiesrepository.org/diecfaori/2

Gutierrez, R., & Giner-Sorolla, R. (2007). Anger, disgust, and presumption of harm as reactions to taboo-breaking behaviors. *Emotion , 7*(4), 853–868.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and Liberals1. *Journal of Applied Social Psychology*, *31*(1), 191–221.

Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The Innate Mind Volume 3: Foundations and the Future.*, *3*(444), 367–391.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? In *Journal of Personality and Social Psychology* (Vol. 65, Issue 4, pp. 613–628). https://doi.org/10.1037/0022-3514.65.4.613

Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. *Thirty-Second Aaai Conference on Artificial*. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16824

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method, and Point*. OUP Oxford.

Hatemi, P. K., Crabtree, C., & Smith, K. B. (2019). Ideology Justifies Morality: Political Beliefs Predict Moral Foundations. *American Journal of Political Science*, *40*, 226.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? The Behavioral and Brain Sciences, 33(2-3), 61–83; discussion 83–135.

Henson, R. G. (1971). Utilitarianism and the Wrongness of Killing. *The Philosophical Review*, *80*(3), 320–337.

Hester, N., Payne, B. K., & Gray, K. (2020). Promiscuous condemnation: People assume ambiguous actions are immoral. *Journal of Experimental Social Psychology*, *86*, 103910.

Hindriks, F. (2014). Normativity in action: how to explain the Knobe effect and its relatives. *Mind & Language*. https://onlinelibrary.wiley.com/doi/abs/10.1111/mila.12041

Hindriks, F., Douven, I., & Singmann, H. (2016). A New Angle on the Knobe Effect: Intentionality Correlates with Blame, not with Praise. *Mind & Language, 31*(2), 204–220.

Hogg, M. A. (2005). All Animals Are Equal but Some Animals Are More Equal than Others: Social Identity

and Marginal Membership. *The Social Outcast: Ostracism, Social Exclusion, Rejection, and Bullying.*, *366*, 243–261.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, *97*(6), 963–976.

Horta, O. (2010). What is speciesism? *Journal of Agricultural & Environmental Ethics*, *23*(3), 243–266.

Icard, T., Cushman, F., & Knobe, J. (2018). On the instrumental value of hypothetical and counterfactual thought. *Mindmodeling.org*. http://mindmodeling.org/cogsci2018/papers/0114/0114.pdf

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and Emotion*, *23*(4), 714–725.

Janoff-Bulman, R., & Carnes, N. C. (2013). Surveying the moral landscape: moral motives and group-based moralities. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *17*(3), 219–236.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*(3), 521–537.

Jenni, K., & Loewenstein, G. (1997). Explaining the Identifiable Victim Effect. *Journal of Risk and Uncertainty*, *14*(3), 235–257.

Kagan, S. (2016). What's wrong with speciesism? (society for applied philosophy annual lecture 2015). *Journal of Applied Philosophy*, *33*(1), 1–21.

Kant, I., & Schneewind, J. B. (2002). *Groundwork for the Metaphysics of Morals*. Yale University Press.

Kean, H., & Howell, P. (2018). *The Routledge Companion to Animal-Human History*. Routledge.

Kim, D.-Y., & Lee, J.-H. (2011). Effects of the BAS and BIS on decision-making in a gambling task. *Personality and Individual Differences*, *50*(7), 1131–1135.

Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J., & Rahwan, I. (2018). A Computational Model of Commonsense Moral Decision Making. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1801.04346

Kivikangas, J. M., Fernández-Castilla, B., Järvelä, S., Ravaja, N., & Lönnqvist, J.-E. (2020). Moral foundations and political orientation: Systematic review and meta-analysis. *Psychological Bulletin*. https://doi.org/10.1037/bul0000308

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015a). Inference of Intention and Permissibility in Moral Decision Making. *CogSci*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.706.3523&rep=rep1&type=pdf

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015b, January 1). *Inference of intention and permissibility in moral decision making | Causality in Cognition Lab*. Causality in Cognition

Lab. http://cicl.stanford.edu/publication/kleiman-weiner2015intention/

Knobe, J. (2003a). Intentional Action and Side Effects in Ordinary Language. *Analysis, 63*(3), 190–194.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*(2), 309–324.

Knobe, J. (2010). Person as scientist, person as moralist. *The Behavioral and Brain Sciences, 33*(4), 315–329; discussion 329–365.

Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences, 7*(1), 67–83.

Kohlberg, L. (1971). Stages of moral development as a basis for moral education. *Moral Education: Interdisciplinary Approaches*, 23–92.

Kohlberg, L. (1973). *Moral development*. McGraw-Hill Films.

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory into Practice, 16*(2), 53–59.

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality, 46*(2), 184–194.

Kollareth, D., & Russell, J. A. (2019). Disgust and the sacred: Do people react to violations of the sacred with the same emotion they react to something putrid? Emotion , 19(1), 37–52.

Kollareth, D., Shirai, M., Helmy, M., & Russell, J. A. (2021). Deconstructing disgust as the emotion of violations of body and soul. Emotion . https://doi.org/10.1037/emo0000886

Kominsky, J. F., & Phillips, J. (2019). Immoral Professors and Malfunctioning Tools: Counterfactual Relevance Accounts Explain the Effect of Norm Violations on Causal Selection. *Cognitive Science, 43*(11), 2386.

Korsgaard, C. (2004). Fellow creatures: Kantian ethics and our duties to animals. *The Tanner Lectures on Human Values*. https://dash.harvard.edu/bitstream/handle/1/3198692/korsgaard_FellowCreatures.pdf

Lagnado, D. A., Gerstenberg, T., & Zultan, R. ’i. (2013). Causal responsibility and counterfactuals. *Cognitive Science, 37*(6), 1036–1073.

Landy, J. F., & Uhlmann, E. L. (2018). Morality is personal. *Atlas of Moral Psychology, 121*. https://books.google.co.uk/books?hl=en&lr=&id=qrk8DwAAQBAJ&oi=fnd&pg=PA121&dq=John+Thain%27s+%2487,000+rug.The+Daily+Beast.+Retrievedfrom&ots=MOgrPdvDNF&sig=NAFrIol8q4oN4kt3NHqV4Y3CMTY

Laurent, S. M., Reich, B. J., & Skorinko, J. L. M. (2019). Reconstructing the side-effect effect: A new way of understanding how moral considerations drive intentionality asymmetries. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0000554

Laurent, S. M., Reich, B. J., & Skorinko, J. L. M. (2020). Understanding Side-Effect Intentionality Asymmetries: Meaning, Morality, or Attitudes and Defaults? *Personality & Social Psychology Bulletin*, 146167220928237.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting Intentionally and the Side-Effect Effect: Theory of Mind and Moral Judgment. *Psychological Science*, *17*(5), 421–427.

Levine, S., & Leslie, A. (2020). *Preschoolers use the means-ends structure of intention to make moral judgments*. https://doi.org/10.31234/osf.io/np9a5

Lin, Z., Yu, J., & Zhu, L. (2019). Norm status, rather than norm type or blameworthiness, results in the side-effect effect. *PsyCh Journal*. https://doi.org/10.1002/pchj.292

Machery, E. (2008). The Folk Concept of Intentional Action: Philosophical and Experimental Issues. *Mind & Language*, *23*(2), 165–189.

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). BayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541.

Malle, B. F. (2021). What the mind is [Review of *What the mind is*]. *Nature Human Behaviour*, *5*(10), 1269–1270.

Malle, B. F., & Knobe, J. (1997). The Folk Concept of Intentionality. In *Journal of Experimental Social Psychology* (Vol. 33, Issue 2, pp. 101–121). https://doi.org/10.1006/jesp.1996.1314

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.

Martin, J., Young, L., & McAuliffe, K. (2019). *The psychology of partner choice*. https://doi.org/10.31234/osf.io/weqhz

Martin, J., Young, L., & McAuliffe, K. (2020). *The impact of group membership on punishment versus partner choice*. https://doi.org/10.31234/osf.io/5qr32

Maxwell, R. (2019, March 5). Why are urban and rural areas so politically divided? *The Washington Post*. https://www.washingtonpost.com/politics/2019/03/05/why-are-urban-rural-areas-so-politically-divided

Michael, J. A., & Szigeti, A. (2019). "The Group Knobe Effect": evidence that people intuitively attribute agency and responsibility to groups. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, *22*(1), 44–61.

Migliore, S., Curcio, G., Mancini, F., & Cappa, S. F. (2014). Counterfactual thinking in moral judgment: an experimental study. *Frontiers in Psychology*, *5*, 451.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152.

Miller, M. R. (2013). Descartes on Animals Revisited. *Journal of Philosophical Research*, *38*, 89–114.

Mill, J. S. (1887). *Utilitarianism*. Willard Small.

Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology*, *93*(1), 1–11.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package "bayesfactor." *URLh Http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor Pdf I (accessed 1006 15)*. ftp://192.218.129.11/pub/CRAN/web/packages/BayesFactor/BayesFactor.pdf

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PloS One*, *14*(8), e0219704.

Muehlhauser, L. (2018, January 15). *2017 Report on Consciousness and Moral Patienthood*. https://www.openphilanthropy.org/2017-report-consciousness-and-moral-patienthood

Nadelhoffer, T. (2004a). On Praise, Side Effects, and Folk Ascriptions of Intentionality. *Journal of Theoretical and Philosophical Psychology*, *24*(2), 196–213.

Nadelhoffer, T. (2004b). The Butler Problem Revisited. *Analysis*, *64*(3), 277–284.

Nakamura, K. (2018). Harming is more intentional than helping because it is more probable: the underlying influence of probability on the Knobe effect. *Journal of Cognitive Psychology* , *30*(2), 129–137.

Nemerever, Z., & Rogers, M. (2021). Measuring the Rural Continuum in Political Science. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, *29*(3), 267–286.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The knobe effect revisited. *Mind & Language*, *22*(4), 346–365.

Noë, R. (2006). Cooperation experiments: coordination through communication versus acting apart together. *Animal Behaviour*, *71*(1), 1–18.

Nozick, R. (1974). Constraints and Animals. *Anarchy, State and Utopia*, 35–42.

Nucci, L. P. (2001). *Education in the Moral Domain*. Cambridge University Press.

Nucci, L. P., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development*, *49*(2), 400–407.

Paprzycka-Hausman, K. (2018). Knowledge of consequences: an explanation of the epistemic side-effect effect. *Synthese*. https://doi.org/10.1007/s11229-018-01973-1

Pearce, D. (2005). The Pinprick Argument. *Utilitarianism Resources*.

Pellizzoni, S., Siegal, M., & Surian, L. (2009). Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology*, *45*(1), 289–295.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), 4649–4654.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, *33*(1), 65–94.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*. https://www.sciencedirect.com/science/article/pii/S1364661319302311

Piaget, J. (1932). *The moral development of the child*. London: Kegan Paul.

Piazza, J., Landy, J. F., Chakroff, A., Young, L., & Wasserman, E. (2018). What disgust does and does not do for moral cognition. *The Moral Psychology of Disgust*, 53–81.

Piazza, J., Sousa, P., Rottman, J., & Syropoulos, S. (2018). Which Appraisals Are Foundational to Moral Judgment? Harm, Injustice, and Beyond. *Social Psychological and Personality Science*, 1948550618801326.

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In *The social psychology of morality: Exploring the causes of good and evil* (Vol. 440, pp. 91–108). American Psychological Association.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychological Science*, *14*(3), 267–272.

Preston, J. L., & Ritter, R. S. (2012). Cleanliness and godliness: Mutual association between two kinds of personal purity. *Journal of Experimental Social Psychology*, *48*(6), 1365–1368.

Preston, J. L., & Shin, F. (2021). Anthropocentric biases in teleological thinking: How nature seems designed for humans. *Journal of Experimental Psychology. General*, *150*(5), 943–955.

Quillien, T. (2020). When do we think that X caused Y? *Cognition*, *205*(104410), 104410.

Quillien, T., & German, T. C. (2021). A simple definition of "intentionally." In *Cognition* (Vol. 214, p. 104806). https://doi.org/10.1016/j.cognition.2021.104806

Ritter, R. S., Preston, J. L., Salomon, E., & Relihan-Johnson, D. (2016). Imagine no religion: Heretical disgust, anger and the symbolic purity of mind. *Cognition & Emotion*, *30*(4), 778–796.

Rollin, B. E. (1989). The unheeded cry: Animal consciousness, animal pain and science. *Studies in Bioethics.*, *308*. https://psycnet.apa.org/fulltext/1989-98197-000.pdf

Royzman, E., Atanasov, P., Landy, J. F., Parks, A., & Gepty, A. (2014). CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the divinity code. Emotion , 14(5), 892–907.

Royzman, E., & Kurzban, R. (2011). Minding the Metaphor: The Elusive Character of Moral Disgust. Emotion Review: Journal of the International Society for Research on Emotion, 3(3), 269–271.

Russell, P. S., & Giner-Sorolla, R. (2011). Moral Anger Is More Flexible Than Moral Disgust. *Social Psychological and Personality Science*, *2*(4), 360–364.

Sabo, J. S., & Giner-Sorolla, R. (2017). Imagining wrong: Fictitious contexts mitigate condemnation of harm more than impurity. *Journal of Experimental Psychology. General, 146*(1), 134–153.

Şahin, M., & Aybek, E. (2019). Jamovi: An easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education*, 670–692.

Schaller, M., & Park, J. H. (2011). The Behavioral Immune System (and Why It Matters). *Current Directions in Psychological Science, 20*(2), 99–103.

Schein, C., & Gray, K. (2015). The Unifying Moral Dyad: Liberals and Conservatives Share the Same Harm-Based Moral Template. *Personality & Social Psychology Bulletin, 41*(8), 1147–1163.

Schein, C., & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc, 22*(1), 32–70.

Schein, C., Ritter, R. S., & Gray, K. (2016). Harm mediates the disgust-immorality link. *Emotion , 16*(6), 862–876.

Schelling, C. T. (1968). The life you save may be your own. *Problems in Public Expenditure*, 127–162.

Schino, G., & Aureli, F. (2017). Reciprocity in group-living animals: partner control versus partner choice. *Biological Reviews of the Cambridge Philosophical Society, 92*(2), 665–672.

Sheikh, S., & Janoff-Bulman, R. (2010). The "Shoulds" and "Should Nots" of Moral Emotions: A Self-Regulatory Perspective on Shame and Guilt. *Personality & Social Psychology Bulletin, 36*(2), 213–224.

Shook, N. J., Thomas, R., & Ford, C. G. (2019). Testing the relation between disgust and general avoidance behavior. *Personality and Individual Differences, 150*, 109457.

Sidgwick, H. (1981). *The Methods of Ethics*. Hackett Publishing.

Singer, P. (1995). *Animal Liberation*. Random House.

Singer, P. (2011). *Practical Ethics*. Cambridge University Press.

Small, D. A., & Loewenstein, G. (2003). Helping a Victim or Helping the Victim: Altruism and Identifiability. *Journal of Risk and Uncertainty, 26*(1), 5–16.

Small, D. A., & Loewenstein, G. (2005). The devil you know: the effects of identifiability on punishment. *Journal of Behavioral Decision Making, 18*(5), 311–318.

Smetana, J. G. (2006). *Social-cognitive domain theory: Consistencies and variations in children's moral and social judgments In Killen M. & Smetana JG (Eds.), Handbook of moral development (pp. 119--154)*. Mahwah, NJ: Lawrence Erlbaum.[Google Scholar].

Smith, C. T., Ratliff, K. A., Redford, L., & Graham, J. (2019). Political ideology predicts attitudes toward moral transgressors. *Journal of Research in Personality, 80*, 23–29.

Sripada, C. S. (2010). The Deep Self Model and asymmetries in folk judgments about intentional action.

Philosophical Studies, *151*(2), 159–176.

Stanley, M. L., Yin, S., & Sinnott-Armstrong, W. (2019). *A reason-based explanation for moral dumbfounding*. http://journal.sjdm.org/18/181220a/jdm181220a.pdf

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, *47*(6), 1249–1254.

Tepe, B., & Aydinli-Karakulak, A. (2019). Beyond harmfulness and impurity: Moral wrongness as a violation of relational motivations. *Journal of Personality and Social Psychology*, *117*(2), 310–337.

Tetlock, P. E. (2003). Thinking the unthinkable: sacred values and taboo cognitions. *Trends in Cognitive Sciences*, *7*(7), 320–324.

Thompson, J. J. (1990). *The Realm of Rights*.

Thomson, R., Yuki, M., Talhelm, T., Schug, J., Kito, M., Ayanian, A. H., Becker, J. C., Becker, M., Chiu, C.-Y., Choi, H.-S., Ferreira, C. M., Fülöp, M., Gul, P., Houghton-Illera, A. M., Joasoo, M., Jong, J., Kavanagh, C. M., Khutkyy, D., Manzi, C., … Visserman, M. L. (2018). Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(29), 7521–7526.

Thornhill, R., & Fincher, C. L. (2014). *The Parasite-Stress Theory of Values and Sociality: Infectious Disease, History and Human Values Worldwide*. Springer.

Tisak, M. S., & Turiel, E. (1984). Children's Conceptions of Moral and Prudential Rules. *Child Development*, *55*(3), 1030–1039.

Toi, M., & Batson, C. D. (1982). More evidence that empathy is a source of altruistic motivation. *Journal of Personality and Social Psychology*, *43*(2), 281–292.

Turiel, E. (1977). Distinct conceptual and developmental domains: social convention and morality. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, *25*, 77–116.

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.

Tybur, J. M., Inbar, Y., Aarøe, L., Barclay, P., Barlow, F. K., de Barra, M., Becker, D. V., Borovoi, L., Choi, I., Choi, J. A., Consedine, N. S., Conway, A., Conway, J. R., Conway, P., Adoric, V. C., Demirci, D. E., Fernández, A. M., Ferreira, D. C. S., Ishii, K., … Žeželj, I. (2016). Parasite stress and pathogen avoidance relate to distinct dimensions of political ideology across 30 nations. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(44), 12408–12413.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *10*(1), 72–81.

Uhlmann, E. L., Zhu, L. [lei], & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry. In *European Journal of Social Psychology* (Vol. 44,

Issue 1, pp. 23–29). https://doi.org/10.1002/ejsp.1987

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334.

van Leeuwen, F., Park, J. H., Koenig, B. L., & Graham, J. (2012). Regional variation in pathogen prevalence predicts endorsement of group-focused moral concerns. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *33*(5), 429–437.

Wagemans, F. M. A., Brandt, M., & Zeelenberg, M. (2017). *Disgust sensitivity and moral judgments of purity: The role of transgression weirdness*. https://psyarxiv.com/c5h42/download?format=pdf

Waleszczyński, A., Obidziński, M., & Rejewska, J. (2019). The Significance of the Relationship between Main Effects and Side Effects for Understanding the Knobe Effect. *Organon F*, *26*(2), 228–248.

Walker, A. C., Turpin, M. H., Fugelsang, J. A., & Bialek, M. (2020). *Better the Two Devils You Know, Than the One You Don't: Predictability Influences Moral Judgment*. https://doi.org/10.31234/osf.io/w4y8f

Waytz, A., Iyer, R., Young, L., Haidt, J., & Graham, J. (2019). Ideological differences in the expanse of the moral circle. *Nature Communications*, *10*(1), 4389.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(43), 11374–11379.

Weisman, K., Legare, C. H., Smith, R. E., Dzokoto, V. A., Aulino, F., Ng, E., Dulin, J. C., Ross-Zehnder, N., Brahinsky, J. D., & Luhrmann, T. M. (2021). Similarities and differences in concepts of mental life among adults and children in five cultures. *Nature Human Behaviour*, *5*(10), 1358–1368.

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, *14*(1), e12562.

Wright, C., & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind & Language*, *24*(1), 24–50.