The relational logic of moral inference

Molly J. Crockett[1]*, Jim A.C. Everett[2], Maureen Gill[1], & Jenifer Z. Siegel[3]

[1]Department of Psychology, Yale University
[2]School of Psychology, University of Kent
[3]Zuckerman Neuroscience Institute, Columbia University

*correspondence to: molly.crockett@yale.edu

To appear in: *Advances in Experimental Social Psychology,* vol. 64

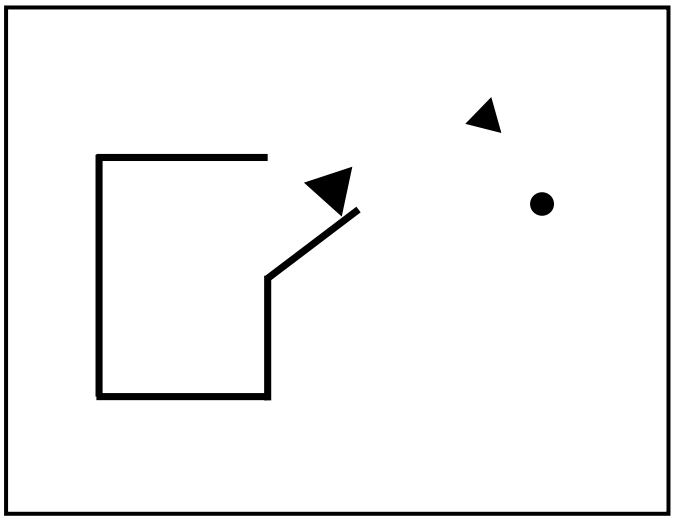
Abstract

How do we make inferences about the moral character of others? Here we review recent work on the cognitive mechanisms of moral inference and impression updating. We show that moral inference follows basic principles of Bayesian inference, but also departs from the standard Bayesian model in ways that may facilitate the maintenance of social relationships. Moral inference is not only sensitive to whether people make moral decisions, but also to features of decisions that reveal their suitability as a relational partner. Together these findings suggest that moral inference follows a relational logic: people form and update moral impressions in ways that are responsive to the demands of ongoing social relationships and particular social roles. We discuss implications of these findings for theories of moral cognition and identify new directions for research on human morality and person perception.

1. Introduction

Heider and Simmel's classic studies of social perception sought to investigate "the processes which are involved in perceiving other individuals, their behavior and their personal qualities" (Heider & Simmel, 1944, p.243). In their experiments, participants viewed a short film depicting movements of two triangles and a circle (Fig. 1).

Figure 1. Heider and Simmel's 1944 experiments presented participants with a 2.5 minute film in which three shapes (a large triangle, a small triangle, and a circle) moved around a larger rectangle in various directions, configurations and speeds. A still frame from this film is depicted below. Figure adapted from Heider & Simmel (1944).



In the first experiment, participants were asked to "write down what happened in the picture." With few exceptions, they wrote rich, detailed narratives that strikingly converged around a few core themes. The participants interpreted the moving shapes as intentional agents, usually humans; they commented on social relationships between the agents, usually perceiving a love triangle; and they evaluated the moral character of the agents, usually identifying a hero and a villain. Here's one example (Heider & Simmel, 1944, p. 247):

> *The first thing we see in this little episode is triangle number-one closing the door of his square. Let's insist that the action of the play is on a two-dimensional surface (not that it makes much difference) and we will undoubtedly start calling the square in which the triangle number-one seems to make his dwelling, a house, which infers three dimensions. But we are not sticking to the theme of our story.*
>
> *Triangle number-one shuts his door (or should we say line) and the two innocent young things walk in. Lovers in the two-dimensional world, no doubt; little triangle number-two and sweet circle. Triangle-one (here-after known as the villain) spies the young love. Ah! ... He opens his door, walks out to see our hero and his sweet. But our hero does not like the interruption (we regret that our actual knowledge of what went on at this particular moment is slightly hazy, I believe we didn't get the exact conversation), he attacks triangle-one rather vigorously (maybe the big bully said some bad word).*

A second experiment asked participants more targeted questions about the agents in the film, (e.g., "What kind of person is the big triangle?"). Again, their responses were remarkably consistent, with 97% of participants describing the large triangle (the "villain") as an aggressive bully, and a plurality describing the small triangle (the "hero") as valiant and brave. A third study showed the film in reverse; again, most participants interpreted the moving shapes as intentional agents with clear motives: a bully chasing a victim, a prisoner escaping a cell, or a parent punishing a child.

Heider and Simmel speculated that their findings reflected a natural tendency to interpret movements in terms of acts of persons with distinctive personality traits: even the sparsest perceptual inputs are imbued with meaning that comes from our experience with the social world. In the decades following this seminal work, much research on social cognition has supported Heider and Simmel's claims. Our minds are built for social inference, and we often go well beyond perceptual inputs when interpreting what we see others do.

Social perception converges along two primary dimensions: whether an agent has good or bad intentions (are they warm, communal, moral?), and whether they are able to carry out those intentions (are they competent, agentic, dominant?) (Fiske et al., 2007; Koch et al., in press). While there remains some disagreement on the exact structure of these dimensions -- e.g., whether morality is distinct from warmth (Brambilla, Sacchi, Rusconi, & Goodwin, 2021), or how the dimensions are applied to stereotypes about social groups (Koch et al., 2020) -- there is widespread consensus that the first thing we notice about others is whether they mean to do us harm or good. This makes sense from an evolutionary standpoint: given the importance of social relationships for survival in ancestral environments, the ability to quickly and accurately detect who is trustworthy provides obvious fitness benefits.

In this chapter, we build on foundational research on social perception to explore the cognitive mechanisms of *moral inference:* how do we learn about the moral character of others from observing their behavior? What are the computational principles that define moral inference and impression updating? What kinds of information do we use to infer the moral character of others? And what consequences do moral inferences have for our own behavior?

We first briefly review social psychological research on the perception of morality, demonstrating that moral impression formation is fast, accurate, and dominates other kinds of social impressions. In particular, this work highlights a negativity bias in moral inference: negative information impacts moral impressions more strongly than positive information (Skowronski & Carlston, 1989). Considering the negativity bias in the context of evolutionary pressures on cooperation highlights a dilemma at the core of moral inference: it needs to accurately identify those who mean us harm, but at the same time, it needs to be sufficiently flexible to account for the fact that people with good intentions sometimes mistakenly cause harm. We then turn to our recent work investigating the cognitive mechanisms of moral inference, which identifies a potential solution to this dilemma and

suggests that moral inference is optimized for social relationships. Further evidence for this claim comes from studies showing that moral inference is not only sensitive to *whether* people make moral decisions, but also the specific types of moral principles they use to guide their decisions that reveal their suitability as relational partners, such as whether they prefer to follow particular moral rules or impartially maximize aggregate welfare. Finally, we consider research suggesting that people make inferences about moral character from information that on the surface appears to have little to do with morality – again, in ways suggesting that moral inference follows a relational logic. We conclude by highlighting the implications of this work for theories of social cognition and identifying future directions for research on morality and person perception.

2. Background

Philosophers and social scientists have long debated the nature of moral character. One important, or perhaps even the central, aspect of moral character is a basic preference for helping over harming others (Gert, 2004). The ability to accurately infer the moral character of others is crucial for successful social functioning. Through recurrent encounters we must learn whether someone's intentions are good or bad, that is, whether they represent an opportunity or a threat. In Section 2, we will summarize past research on moral inference and identify a puzzle it raises: negative information dominates our moral impressions of others, but maintaining cooperative social relationships often requires forgiving others for past transgressions. We'll consider how this puzzle might be resolved in Section 3.

2.1. Morality dominates person perception

Moral traits signal the extent to which an individual cares about another's wellbeing. From an evolutionary perspective, learning whether someone has good or bad intentions (i.e., whether they mean to help or harm us) is more important for survival than whether they can carry out those

intentions (Fiske et al., 2007). Accordingly, Fiske et al. (2007) have argued that moral information is sought out first to anticipate whether someone represents an opportunity or a threat, and only then is competence evaluated to determine *how* valuable or threatening they may be. Supporting this view, there is evidence that people are more interested in acquiring information indicative about others' morality than their competence (Brambilla et al., 2011; De Bruin & van Lange, 2000; Wojciszke et al., 1998).

A large body of evidence shows that we infer moral character rapidly and effortlessly (Brambilla et al., 2021; Engell et al., 2007; Goodwin et al., 2014; Todorov et al., 2008; Todorov & Oh, in press). In adults, impressions of trustworthiness stabilize after merely 100 milliseconds of exposure to novel faces (Willis & Todorov, 2006), and these impressions influence subsequent moral evaluations even when facial information does not reach conscious awareness (Todorov et al., 2009). People readily infer moral character based on a single piece of behavioral information (Inbar et al., 2012; Mende-Siedlecki et al., 2012; Todorov & Uleman, 2003), and moral traits are processed more rapidly than traits indicative of competence (Willis & Todorov, 2006; Ybarra et al., 2001). Such rapid assessments of moral character are likely to be important for identifying the immediate goals of an actor and predicting what they will do in the future.

Once acquired, information relevant to assessing moral character dominates other types of information in person perception. Many studies reveal that our overall evaluations of others depend to a higher degree on moral traits (e.g., trustworthiness, kindness) than non-moral traits (e.g., intelligence, courageousness) (Wojciszke, 2005; Wojciszke et al., 1998). When asked to judge a familiar or novel other, individuals' impressions are more strongly predicted by the target's moral traits than traits indicative of competence or sociability (Goodwin et al., 2014).

Given the primacy of moral information in person perception, people also care strongly about *appearing* moral to others. Across different types of relationships (e.g., family members, employees)

trustworthiness is rated as the most desirable trait for an ideal person to possess, whereas other traits, such as intelligence and conscientiousness, are only valued to the extent that they are relevant to the nature of the relationship (Cottrell et al., 2007). People are also more concerned about their group appearing moral than appearing competent (Leach et al., 2007), and will even go to surprising lengths to preserve their individual reputations. In one study, more than half of participants reported they would rather die than have a reputation as a child-molester, and about a third of participants chose to submerge their hands in a bucket of live worms instead of having an email sent out that described them as racist (A. J. Vonasch et al., 2018). Indeed, evolutionary thinkers have explained the origins of morality and cooperation through the adaptive value of being identified by others as a morally good person (e.g. Alexander, 1987; Baumard et al., 2013; Nowak & Sigmund, 2005; Roberts, 1998). It has even been suggested that our very brain size and capacity for complex language is attributable to our need to appraise and share information – i.e., gossip – about the moral character of others (Dunbar, 2004). Thus, a sensitivity to information about moral character seems to be a defining feature of human social cognition.

2.2. Moral impression updating and the negativity bias

Moral impressions not only help us predict others' behavior, but also guide our own behavior; people will approach those whom they infer to have a good moral character and avoid those whom they infer to have a bad moral character (Brambilla et al., 2013; Pagliaro et al., 2013). Our ability to assess others' morality from social cues helps us reap benefits from social interactions and avoid exploitation. Even implicit social cues, such as the perceived trustworthiness of a face, impact subsequent social decisions in behavioral economic games. For instance, people invest significantly more money with strangers represented by faces rated as more trustworthy, despite no objective relationship between

appearance and the return on investment (FeldmanHall et al., 2018; Stanley et al., 2011; van 't Wout & Sanfey, 2008).

Building and maintaining accurate representations of others' morality is important because inaccurately inferring malintent in kind individuals can lead to missed opportunities, while inaccurately inferring good intent in self-serving individuals exposes us to manipulation and harm. To maintain these representations, we must also continually and flexibly update our impressions of others. Consider a time when you drastically changed your mind about someone. Perhaps a romantic partner cheated on you, or you were pleasantly surprised by the kindness of a coworker with a reputation for being selfish. Changing your impression not only served the purpose of updating your expectations about their future actions but it allowed you to adapt your own decisions given those new expectations. For example, learning that your neighbor has gossiped about another neighbor on your street might not only motivate you to update your impression of their trustworthiness but also teach you to behave more cautiously around that person.

A growing body of research on impression updating reveals how our moral impressions of others are updated in light of new evidence (Cone & Ferguson, 2015; Mende-Siedlecki et al., 2012; Reeder & Coovert, 1986; Skowronski & Carlston, 1992). To study impression updating, researchers have often employed lists of traits (e.g., energetic, persuasive, cold) or phrases describing a target's social behaviors (e.g., "volunteered to stay late to help a coworker" and "kicked a stray cat to get it to leave his yard"). Participants are asked to provide an impression of the target's character after being presented with a subset of the list, and then asked to provide another impression rating after being presented with another subset that is either congruent or incongruent in valence with the earlier behaviors. Impression updating is then quantified as the difference between the initial impression rating and subsequent impression rating. Work using these paradigms has made great strides towards our understanding of how impressions are updated. For example, Asch (1946) proposed that the overall impression of an

individual is not equal to the algebraic sum of its components. That is, each piece of information does not hold an equal weight towards updating an impression (Anderson, 1965).

Consistent with this notion, many studies have demonstrated that immoral actions carry greater weight towards updating an impression than moral actions (Rozin & Royzman, 2001; Skowronski & Carlston, 1989). This is commonly known as the *negativity bias* in impression formation. In an early demonstration of the negativity bias (Reeder & Coovert, 1986), participants read a subset of either highly moral or highly immoral behaviors describing a target individual (e.g., "Donated time to take blind children to the park" and "Hit a child for no reason"). Next, participants rated their impression of the target from 1 (*highly immoral*) to 9 (*highly moral*). Lastly, participants read a final behavior in which the valence was incongruent with the initial subset of behaviors and provided a new impression based on the combined information. The authors found that impressions were more heavily updated when immoral information followed moral information than the other way around. This study, along with others (Briscoe et al., 1967; Cone & Ferguson, 2015; Dibbets, Pauline et al., 2012; Freedman & Steinbruner, 1964; Martijn et al., 1992; Riskey & Birnbaum, 1974), suggest that initial positive character impressions are more amenable to updating when presented with disconfirming evidence than negative character impressions. What might explain the emergence of a negativity bias in updating moral impressions?

One influential hypothesis posits that people consider negative moral information to be more *diagnostic* of someone's character than positive moral information (Skowronski & Carlston, 1987, 1989). Not all actions are equally informative about a person's character (Fiske, 1980). That is, there is a probabilistic relationship between actions and their causes; e.g., selfish people sometimes behave generously, but generous people rarely behave selfishly. Hence, we infer moral character more strongly from immoral actions because moral actions may be less informative for predicting others' behavior. This heuristic may be useful for building expectations about the probability and magnitude of possible

harms. Knowing the line someone will not cross, morally speaking, enables us to more easily and confidently predict how they will behave across a larger number of situations. This is consistent with research showing that people are more confident that someone who behaves selfishly will continue to behave selfishly than someone who behaves generously will continue to behave generously (Martijn et al., 1992).

Another possibility is that the negativity bias in moral impression formation is simply one manifestation of a broader, domain-general phenomenon: namely, that negative stimuli are psychologically more impactful than positive stimuli (Baumeister et al., 2001; Rozin & Royzman, 2001; Unkelbach et al., 2020). However, studies of impression formation on the dimension of competence do not support this view. Here, behaviors that convey a high level of competence carry more weight than incompetent behaviors (Skowronski & Carlston, 1987). The diagnosticity hypothesis can explain this overall pattern (negativity bias for moral inference, positivity bias for competence inference) because highly competent individuals sometimes behave incompetently, but highly incompetent individuals only rarely behave competently (Reeder & Brewer, 1979; Skowronski & Carlston, 1989).

2.3. Cooperation and forgiveness in an uncertain world

A negativity bias in social evaluations is theorized to be adaptive in an evolutionary sense (Cacioppo et al., 1997; Vaish et al., 2008): those who are “better safe than sorry” when inferring moral character should be more likely to avoid exploitation. However, forming rigid negative impressions of others carries its own costs in environments where people sometimes make mistakes. Erroneously inferring bad character can lead people to prematurely terminate valuable relationships and thereby miss out on the potential benefits of future cooperative interactions (R. M. Axelrod, 2006; Johnson et al., 2013; McCullough, 2008; Molander, 1985). Thus, successfully navigating social life requires strategies for maintaining social relationships even when others behave inconsistently and sometimes behave badly.

How can cooperation be maintained in an uncertain world? One possibility is to respond to bad behavior with probabilistic cooperation (Nowak & Sigmund, 1992). Evolutionary models show that this strategy, called "generous tit-for-tat", outcompetes strategies that summarily end cooperative relationships following a single betrayal (Fudenberg et al., 2012; Wu & Axelrod, 1995). In line with these models, there is also evidence that humans use a generous tit-for-tat strategy when playing repeated prisoner's dilemmas where others' intended behaviors are implemented with noise (Fudenberg et al., 2012). Thus, it appears that the most successful strategies for repeated social interactions in an uncertain world are ones that contain an element of forgiveness and leniency (Fudenberg et al., 2012; Rand et al., 2009).

Forgiveness can be understood as the positive amendment of thoughts, behaviors, or emotions towards a transgressor (McCullough, 2000; McCullough et al., 2000; Snyder & Lopez, 2001). Forgiveness is not only abundant in social interactions but also important for an individual's mental and physical health (T. W. Baskin & Enright, 2004; Worthington & Scherer, 2004). Research indicates that the propensity to forgive others is related to fewer symptoms of depression, decreased anxiety, and lower blood pressure (R. P. Brown, 2003; Krause & Ellison, 2003; Maltby et al., 2001; Witvliet et al., 2001). Difficulty forgiving others' minor transgressions is commonly observed in mental illnesses characterized by interpersonal dysfunction (Barnow et al., 2009; Unoka et al., 2009). Patients with Borderline Personality Disorder (BPD) often hold grudges and write people off following seemingly insignificant slights (Sansone et al., 2013; Thielmann et al., 2014). Chronic unforgiveness is therefore believed to be a key component leading to abnormal social cognition and behavior in BPD patients (Gartner, 1988; Holm et al., 2009; Sansone et al., 2013), leading to the integration of forgiveness skills in standard treatments (Sandage et al., 2015).

2.4. Reconsidering the negativity bias

The ability to forgive others may be a necessary component for healthy social functioning. But how do we make sense of evidence for a strong propensity to forgive alongside research on the negativity bias, which suggests bad moral impressions are especially resistant to updating? One possibility is that the negativity bias may not operate in real life the way it does in psychological experiments. The bulk of evidence for the negativity bias comes from paradigms that may not resemble the kind of social information we typically encounter in everyday social interactions. In many of these studies, impression formation was examined using narrative descriptions of extreme and rare behaviors. For instance, Reeder and Coovert (Reeder & Coovert, 1986) studied impressions from negative behaviors such as "Stole money from a charity fund" and "Hit a child for no reason", versus positive behaviors such as "Rescued a family from a burning house" and "Donated time to take blind children to the playground". Not only are such behaviors rare to encounter in real life, but it is also highly unlikely that *the same person* would engage in both types of positive and negative behaviors. Results from studies like this may therefore not generalize to moral impression updating in real life, where most of us are unlikely to routinely interact with people who have rescued families from a burning house or stolen from charity.

Interestingly, when impression formation is examined using more minor transgressions and everyday acts of kindness (e.g., "He gave out toys at the children's hospital" and "He told a colleague in public that she should lose weight"), reports of the negativity bias are less consistent (Carr & Walther, 2014). Skowronski and Carlston (Skowronski & Carlston, 1992) found that the difficulty in revising an initial bad impression was directly related to the extremity of behaviors; less extreme behaviors were easier to update in light of disconfirming evidence than more extreme behaviors. Consistently, Wojcizske et al. (1993) found that the negativity bias was either weak or non-existent when only moderately good and bad behaviors were used. Understanding moral inference from moderate behaviors is important because more moderate behaviors comprise the vast majority of what we

experience on a daily basis, and our success as a social species suggests we are able to accurately form moral impressions even in the absence of highly diagnostic information.

In a similar vein, recent work suggests that the negativity bias in impression updating can be explained by perceptions of how rare immoral behaviors are, relative to moral ones, and report no evidence for a negativity bias after controlling for the perceived frequency of behaviors (Mende-Siedlecki et al., 2013). Most studies on moral impression formation have not explicitly controlled for the perceived frequency of positive and negative behaviors. Thus, it is unclear whether valence asymmetries in impression updating observed in the past literature were due to differences in perceived frequency of behaviors, which we would expect to have downstream effects on diagnosticity. If negative behaviors used in these studies were in fact more rare than positive behaviors (and by extension, more diagnostic of character), participants may have formed more certain impressions about the bad characters than the good characters, which could explain why bad impressions in these studies were less amenable to updating. This leaves open the question of whether people actually learn differently about people inferred to be more vs. less moral, when their actions are equally diagnostic of their underlying character. Previous work has been unable to address this question because these studies have typically examined how people update impressions in response to vignettes featuring definitively (im)moral behaviors. This methodology makes it very difficult, if not impossible, to match the frequency and diagnosticity of moral vs. immoral behaviors because perceptions of vignettes are inherently subjective.

Another limitation of previous work on moral impression updating is that most studies employ relatively few behaviors and only a single 'update'. That is, participants often rate their impression at only two time points; first after being presented with one description of behavior, and again after being presented with another set of behaviors that are incongruent in valence. However, evaluations of others rely on probabilistic inferences made about people's actions, intentions and motivations, which unfold over time and are influenced by multidimensional factors including past experiences. Standard vignette-

based paradigms employed by much of the work on impression formation are therefore limited in their ability to capture the complex dynamics of the moral inference process.

2.5. Summary

Moral information dominates person perception, and there is substantial experimental evidence for a negativity bias in moral impression updating. Nevertheless, there are several important questions to resolve when considering how moral inference operates in everyday life. First, because past work on impression formation has mainly focused on extreme (and therefore highly diagnostic) behaviors, it is still relatively unknown how people form moral impressions from more moderate behaviors. Second, it is not well understood how moral impressions dynamically change over time. And finally, although a negativity bias helps us avoid exploitation, it can also be costly: attributing bad character from rare or minor transgressions leaves little room for people to make mistakes and can impede the development of stable relationships necessary for healthy social functioning. Responding to immoral acts with leniency and forgiveness may enable the development of successful relationships despite occasional harms. Although evolutionary and economic models provide descriptive accounts of these behaviors (Fudenberg et al., 2012; Grim, 1995; Nowak & Sigmund, 1992; Rand et al., 2009), the cognitive mechanisms that enable them are not well understood. These questions will be addressed in the next section.

3. Computational principles of moral inference

How do people dynamically form and update moral impressions from moderately helpful and harmful behaviors? In this section, we take a computational approach to develop a hypothesis that moral inference is optimized for relationship maintenance. Computational approaches can sidestep some of the limitations of past research and reveal new insights into the dynamic cognitive mechanisms

of moral inference and impression updating in healthy people and in clinical populations with disturbed social relationships.

The cognitive scientist David Marr famously argued that to properly understand any information processing system, one must interrogate the system at three separate levels of analysis (Marr, 1982). The *computational level* identifies the goals of the system: what kinds of problems does moral inference solve, and what are the normative solutions to these problems? The *algorithmic level* specifies formally how inputs are transformed into outputs: how are features of moral actions (e.g., harms to others, benefits to self) transformed into global impressions of moral character? Finally, the *implementation level* describes how the algorithms are implemented in neural hardware.

Here, we focus primarily on the computational level of analysis. As a starting point, we consider Bayesian models of learning and inference under uncertainty, which provide a normative framework for building models of our environment under conditions of uncertainty. In other words, Bayesian models describe how we should optimally form beliefs when confronted with noisy data. Foundational work in the cognitive science of learning and decision-making has primarily considered the computational goal of accuracy: learners should seek to form accurate beliefs. Our recent work suggests an additional computational goal for *moral* inference: learners should seek to form accurate beliefs, but also should update those beliefs in ways that facilitate the development and maintenance of social relationships.

3.1. Bayesian models of learning and inference

Recent studies in cognitive science have sought to investigate the cognitive mechanisms through which people infer other's hidden preferences, intentions, and desires over time (Aksoy & Weesie, 2014; Diaconescu et al., 2014; Jern & Kemp, 2015). By measuring the behaviors of individual learners and then fitting descriptive models to learners' behavior, researchers can determine the extent to which behavior conforms to the predictions of the Bayesian ideal. These methods have advanced our understanding of

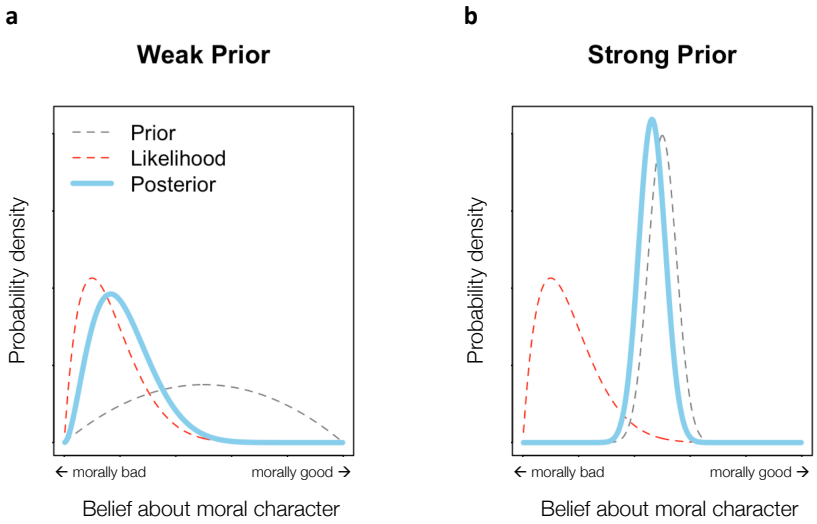
human cognition by allowing researchers to quantify trial by trial dynamics of learning and decision making to make precise predictions about behavior (Charpentier & O'Doherty, 2018; Hackel & Amodio, 2018; Lockwood & Klein-Flugge, 2020).

Bayesian inference models describe how an "ideal" observer should use accumulating evidence to optimally update beliefs about hidden states (e.g., preferences, desires, intentions) in the form of probability distributions. Bayesian models describe how *prior beliefs* (i.e., expectations) are updated with information about the *likelihood* of evidence, to arrive at *posterior beliefs* (i.e., current estimate of the hidden state; **Fig. 2**). In social interactions, prior beliefs about others' character traits are used to make predictions about their future behavior. To illustrate, imagine you are standing behind your co-worker, George, in a coffee shop. George hands the cashier a $10 bill to pay for his coffee and you notice that the cashier accidentally hands him back a $50 bill instead of a $5 bill for change. You probably expect that George will take a series of actions: speak to the cashier, inform them of the mistake, and return the change.

Now, imagine George folds the bill and places it in his pocket. Consider a scenario where George had just started working with you. Because your experience with George is limited, your prior belief about his moral character is weak (**Fig. 2a**) – you are not confident about whether he is honest or dishonest. Consequently, the new evidence (pocketing the excess change) makes a strong impression, and you ultimately come to believe George is dishonest. Now consider a scenario where you have worked with George for a long time and are highly confident that he is reasonably honest – perhaps not a saint, but pretty unlikely to intentionally steal $45 (**Fig. 2b**). In this case, George's behavior does not have much impact, and your posterior belief is not far from your prior. Instead, you may search for alternative explanations for George's behavior -- for example, you may hypothesize that George didn't notice the mistaken change – to help maintain your positive impression. These divergent outcomes illustrate the

impact of uncertainty in our prior beliefs on impression updating, a measure which has more often than not been overlooked in the impression formation literature.

Figure 2. Bayesian updating of beliefs about moral character. (a) A scenario where an observer has a weak prior belief about a target's moral character (dashed gray line). Upon observing a dishonest behavior (dashed red line), the posterior belief (solid blue line) is updated in line with the evidence. **(b)** A scenario where an observer has a strong prior belief about a target's moral character. The strong prior limits the extent of belief updating, such that even after observing a dishonest behavior, the posterior belief has not shifted much relative to the prior.



In Bayesian inference, the *variance* of the prior and posterior probability distributions captures how uncertain we are about our initial and updated beliefs. Bayesian inference suggests how we might optimally incorporate this type of uncertainty, often called estimation uncertainty, towards updating our impressions of others. According to this framework, uncertainty should suppress top-down expectation guided processes (prior expectations), but boost bottom-up sensory induced signals to promote learning about the uncertain state (Yu & Dayan, 2003). To this end, estimation uncertainty is related to the rate

of updating our beliefs in the face of new information (i.e., sensory signals), where high levels of uncertainty call for faster updating. This relationship between estimation uncertainty and the learning rate is normative: the more uncertain our beliefs, the more we should incorporate new evidence into our beliefs. Accordingly, a weak prior belief that George has honest preferences should motivate belief updating when presented with evidence that does not support this belief. In this way, even if initial beliefs are unreliable, they might still serve as a basis for learning about others by motivating more accurate models of a person's character via enhanced information seeking and updating from evidence.

Ample research shows that in both social and non-social settings, human learning broadly conforms to the predictions of Bayesian models: all kinds of human beliefs are updated in proportion to their uncertainties (Griffiths et al., 2008). Whether people actually perform Bayesian statistics in their heads when learning under conditions under uncertainty is a question that remains unresolved. Nevertheless, Bayesian models serve as a useful benchmark against which to measure human learning because they specify precisely how to optimally make predictions in an uncertain world – a problem humans continually face, especially during social interactions.

One useful application of Bayesian learning models is a program of research demonstrating how learning responds to physiological states such as arousal. This work shows that being in an aroused state accelerates belief updating by increasing the susceptibility of new information on existing beliefs. Nassar et al. (2012) directly examined the relationships between arousal and non-social perceptual learning in a predictive-inference task where participants inferred the mean of a gaussian distribution that changed stochastically over time. The authors found that pupil diameter, which is believed to be a reliable indicator of state arousal (Eldar et al., 2013; McGinley et al., 2015; Reimer et al., 2016), was positively related to uncertainty about the inferred mean and subsequent belief updating, as estimated from a Bayesian learning model (Nassar et al., 2012). Importantly, the authors found that a task-independent manipulation of arousal increased belief uncertainty and belief updating by increasing the susceptibility

of new information on existing beliefs. Recent work using a threat-of-shock manipulation suggests how this process might unfold. Threat-induced arousal was shown to increase sensitivity to new information by augmenting the influence of prediction errors via amplified cortical excitability while dampening feedback from prefrontal cortices to reduce reliance on prior expectations (Cornwell et al., 2017). Such an arousal-linked learning system may be especially advantageous in the face of impending threats, where the ability to rapidly detect and respond to one's environment is essential for avoiding harm.

The fact that threat-induced arousal accelerates belief updating suggests a possible solution to the puzzle identified in Section 2: people might form more uncertain beliefs about threatening social stimuli (such as people behaving badly), and therefore may be able to rapidly update those beliefs if they turn out to be mistaken. Because social threats are also emotionally and physiologically arousing (Fouragnan, 2013; Noordewier et al., 2019; Öhman, 1986; Roelofs et al., 2010), people may be especially uncertain about immoral agents and consequently augment the relative influence of newly arriving over historical information on existing beliefs. In other words, beliefs about bad agents may be updated more rapidly, because uncertainty associated with potential threats promotes learning from new information. Conversely, diminished vigilance may suppress updating from new information to favor rapidly developed (prior) assessments when evaluating the choices of good agents. This suggests a theoretical model where inferences about moral character impact the relative influence of historical and new information on existing beliefs in impression formation. More uncertain, flexible beliefs may help maintain relationships in the wake of immoral behaviors by enabling them to update rapidly in light of new information. Consequently, examining the computations underlying moral inference may elucidate the mechanisms through which people can update negative impressions despite the potency of negative information in impression formation. Given the importance of cooperation and forgiveness in maintaining healthy relationships, these processes may be crucial for understanding populations displaying dysfunctional social cognition and behavior.

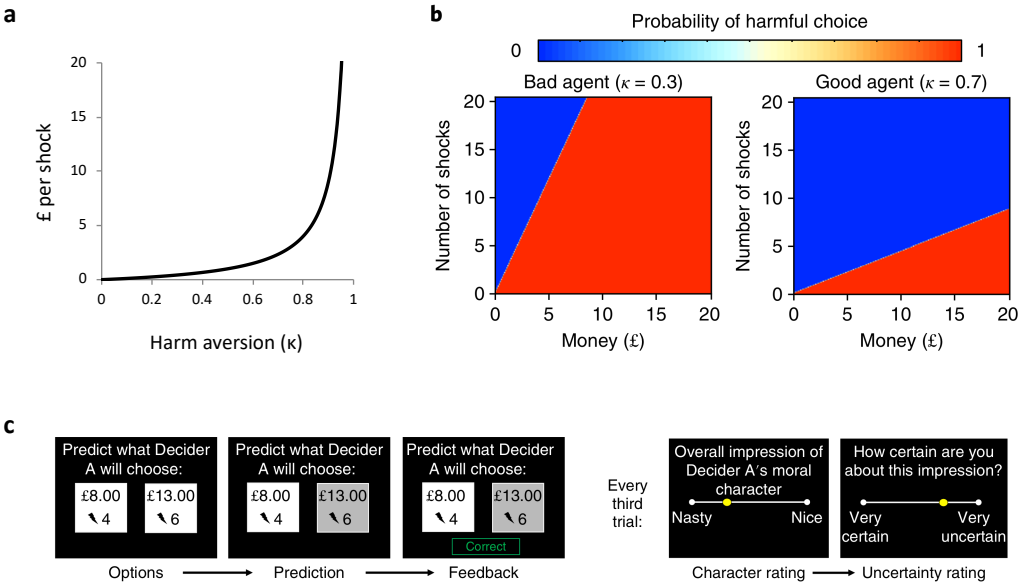
3.2. Investigating the computational mechanisms of moral inference

We conducted a series of experiments to test the hypothesis that, when people form beliefs about others, they form more *uncertain* impressions about harmful individuals than helpful individuals (Siegel et al., 2018). This makes negative impressions more *volatile* (i.e., amenable to updating), while positive impressions would remain more stable. Such an asymmetry would provide a possible solution for maintaining social relationships when others sometimes behave badly: more negative impressions can be easily revised if they turn out to be mistaken, while more positive beliefs are resilient and can remain stable in the face of minor transgressions.

To test this hypothesis, we developed a paradigm that enabled us to control for a number of factors known to impact impression updating. First, we know that moral impressions are highly dependent on prior social experience, which is difficult to control. To account for this, we measured moral inference in a setting where people observe morally relevant actions that they have not encountered before and therefore would not have strong prior expectations about. We exploited a paradigm we developed to study moral decision-making in the lab, where people decide whether to inflict moderately painful electric shocks on another person in exchange for money (Crockett et al., 2014). Decisions in this paradigm are well-described by a model that includes a "harm aversion" parameter, κ, which quantifies the subjective cost of harming the victim as an exchange rate between money and pain and ranges from 0 (profit maximizing) to 1 (pain minimizing) (**Fig. 3a**). Because ethical systems universally prohibit harming others for personal gain (Gert, 2004), and because harm aversion correlates with other morally relevant traits like empathy and psychopathy (Crockett et al., 2014), we can use harm aversion as a proxy for moral character in this paradigm. Of course, the decision to trade money for shocks is unlikely to be encountered in everyday life. However, in the context of studying moral inference, this is a feature (rather than a bug) because participants are unlikely to hold strong

prior beliefs about typical levels of harm aversion in the population. Decisions to seek or avoid profit from inflicting (mild) pain on another person are also more moderate than the types of behaviors typically studied in impression formation research. This allowed us to study how people form moral impressions from information that is not extreme or definitive, which is closer to the information that people use when forming impressions in their daily lives.

Figure 3. Moral inference task. (a) In the task, two "agents" make decisions about whether to inflict electric shocks on another person in exchange for money. Preferences are captured by a "harm aversion" parameter, here plotted as a function of how much money (in British pounds) an agent requires to deliver an additional shock to the victim. **(b)** Participants learned about two agents who differed in their harm aversion; the "good" agent required substantially more money per shock than the "bad" agent. **(c)** On each trial, participants were shown the agent's choice options; predicted what the agent would choose; and shown the agent's actual choice. Every third trial, participants provided a subjective rating of the agent's overall moral character, and a subjective rating of the uncertainty of their impression. Panels b-c reprinted with permission from Siegel et al. (2018).



In our moral inference task, we presented participants with two 'agents' who faced a series of decisions trading off different amounts of money for oneself and pain to a victim. One agent (the 'bad' agent) had a low level of harm aversion (κ = 0.3), requiring only a profit of $0.43 to deliver each shock to the victim. Another agent (the 'good' agent) was substantially more harm averse (κ = 0.7), requiring $2.40 per shock (**Fig. 3b**). Of course, the terms 'bad' and 'good' are subjective and relative; here our main interest was to probe whether the dynamics of moral impression formation differed when learning about agents who differed in their levels of harm aversion. Participants observed sequences of 50 choices made by each agent (presented in randomized order). On each trial (**Fig. 3c**), participants predicted the choice the agent would make and immediately received feedback about whether their prediction was correct. Every third trial, participants made a global impression of the agent's moral character (on a scale from 'nasty' to 'nice' or 'bad' to 'good') and also rated the uncertainty of their impression (on a scale from 'very certain' to 'very uncertain'). After learning about both agents, participants were invited to entrust money with each agent in a one-shot trust game.

The task design therefore allowed us to measure three different aspects of moral inference. By observing the trajectories of the impression ratings, we captured participants' evolving *subjective impressions* of the agents' moral character, and the uncertainties of those impressions. A Bayesian reinforcement learning model was fit to participants' trial-wise behavioral predictions of the agents' choices to capture the influence of historical information and newly arriving information on evolving *beliefs about the agents' harm aversion*. And finally, by observing the money entrusted in each agent, we could measure the *behavioral consequences* of moral inference.

Instead of presenting participants with positive or negative behaviors in hypothetical scenarios (as in past research on moral impression formation), our paradigm operationalized moral character in terms of an exchange rate between money and pain, allowing us to tightly control how informative agents' behavior was with regard to their underlying preferences (i.e., the harm aversion parameter). This allowed us to investigate impression formation from positive and negative behaviors that were matched in diagnosticity, sidestepping the challenges inherent in attempting to match positive and negative stimuli using vignettes. Over the course of learning, we precisely matched the trial sequences with respect to how much information was provided about each agent's harm aversion. In this way, we could ensure that the statistics of the environment did not advantage learning about either the good or bad agent, and this symmetry was confirmed by the fact that an ideal Bayesian observer (i.e., a simulated learner that performs the task as accurately as possible) learned identically about the good and bad agents. Because of this design feature, if we observed an asymmetry in learning about bad versus good agents, we could confidently infer these this is not due to an asymmetry in the information we provide to participants. And by comparing participants' actual learning behavior against the benchmark of the ideal Bayesian observer, we were able to discover additional computational goals that may be unique to *moral* inference.

3.3. Beliefs about 'bad' agents are more uncertain and volatile than beliefs about 'good' agents

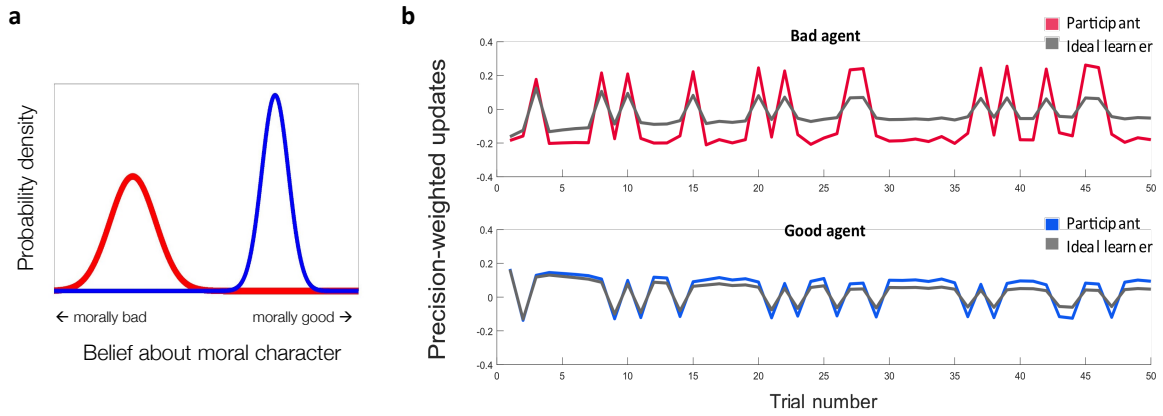
Siegel et al. (2018) used the moral inference task to investigate the computational mechanisms of moral inference in a series of lab and online experiments. In each study, participants predicted the choices of a good and bad agent, and these data were modelled with a Bayesian reinforcement learning model that generated a global estimate of belief volatility (parameter, ω) that captures how rapidly beliefs evolve over time. Belief volatility is set in log space and is monotonically related to belief uncertainty as measured by belief variance, σ (i.e., more uncertain beliefs are more volatile (Mathys et al., 2011). Bayesian reinforcement learning models capture the fact that when beliefs are uncertain, they are more sensitive to updating from new information; thus the rate of learning can change in tandem with belief uncertainty, in contrast to simpler reinforcement learning models (such as the Rescorla-Wagner model) where learning rates are fixed and do not change over time. Across all of our moral inference studies, Bayesian models described participants' behavior better than simpler Rescorla-Wagner learning models, demonstrating that social inference is dynamic and highly sensitive to the uncertainty of beliefs.

Because we operationalized moral character in quantitative terms, we were actually able to make claims about the accuracy of moral inference, because there is a "ground truth" about the agents' moral preferences that is discoverable (in contrast to hypothetical scenarios, where moral inference is subjective). Examining the learning trajectories revealed that participants were quickly able to develop accurate beliefs about agents' harm aversion; the model's final estimates of participants' beliefs about each agent's κ closely resembled the agent's true κ and significantly differed from one another. Subjective character ratings showed a similar pattern; participants inferred positive and negative moral character for agents with high and low harm aversion, respectively. Participants were able to use these

divergent beliefs and impressions to adaptively tune their trust behavior, entrusting the good agent with about twice as much money than the bad agent in a trust game.

In line with our predictions, we also observed a striking asymmetry in the uncertainty and volatility of beliefs about good and bad agents. Across studies, beliefs about the harm aversion and impressions of the moral character of bad agents were consistently more uncertain and volatile than beliefs about good agents. Effectively, this means that prediction errors carried more weight when updating beliefs about the bad agents' harm aversion than the good agents' harm aversion (**Fig. 4**). This asymmetry was observed regardless of the labels participants used to rate the agents' character (e.g., "nasty/nice" versus "good/bad"), whether the studies were conducted in the lab versus online, and whether the agents implemented their preferences with high or low randomness. Notably, this asymmetry is not present in an ideal Bayesian observer model that seeks only to form accurate beliefs. This departure of human moral inference from the Bayesian ideal suggests that successful moral inference might require solving for additional goals beyond accuracy. In other words, the patterns of belief updating that are "ideal" or "optimal" in the traditional Bayesian sense may not be "ideal" or "optimal" in the context of moral inference.

Figure 4. Beliefs about morally bad agents are more uncertain and volatile than beliefs about morally good agents. (a) Schematic illustration of the overall pattern of findings. **(b)** Prediction errors were ascribed greater weights when updating beliefs about the bad agents' harm aversion than the good agents' harm aversion. Asymmetric belief updating was significantly greater in human participants compared to an ideal Bayesian learner.

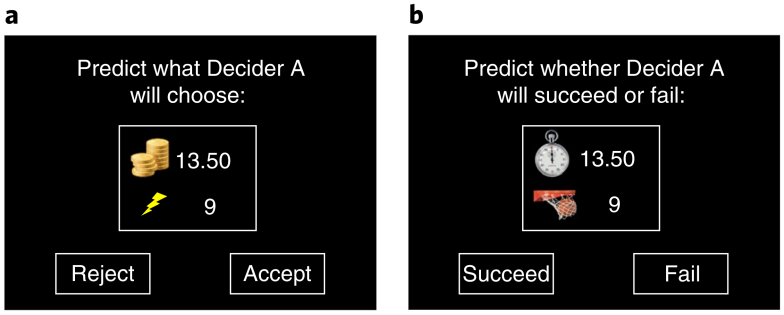


One potential explanation for why beliefs are more uncertain about morally worse agents is that people may hold strong prior expectations that most people will behave morally (Brañas-Garza et al., 2017; Levine et al., 2018). This would make the behavior of bad agents more surprising, resulting in more uncertain beliefs in accordance with Bayes' rule. We tested this possibility by asking a separate group of participants to predict how "most people" would make tradeoffs between money for themselves and pain for others. These data showed that participants' expectations for such decisions were equidistant from the behaviors of the morally good and bad agents, which meant that in the context of our task, the bad agent's behavior was no more (or less) surprising than the good agent's behavior. Additionally, in some studies we measured participants' prior beliefs about the moral character of the agents before observing any of their decisions. We found no relationship between these prior beliefs and the extent of the asymmetry in belief uncertainty for bad versus good agents, suggesting the asymmetry cannot be fully explained by prior expectations about others' morality in general.

Is the asymmetry we observed specific to learning about positive versus negative *moral* traits, or can it be observed when learning about positive versus negative traits more generally? We tested this in a study where we examined whether the asymmetry in learning about bad compared to good agents extended to learning about a trait unrelated to morality. Participants were randomized to either a

morality condition or a competence condition. In the morality condition (**Fig. 5a**), participants predicted whether agents (high or low harm aversion) would accept or reject "offers" to inflict a certain amount of pain for a certain amount of money. In the competence condition (**Fig. 5b**), participants predicted whether agents (high or low skill) would succeed or fail at scoring a certain number of points in a certain amount of time in a series of basketball games. The information provided to participants about the agents' morality and competence was matched perfectly between conditions. Thus, any differences in learning observed between the two conditions could only be attributable to differences in the framing of the tasks.

Figure 5. Inferring morality and competence. The morality condition **(a)** and competence condition **(b)** were perfectly matched in terms of the information provided about the hidden variable participants were asked to infer. Figure reprinted with permission from Siegel et al. (2018).



In this study, participants formed accurate beliefs about the agents' harm aversion in the morality condition, and about the agents' skill levels in the competence condition. However, the uncertainty and volatility of beliefs differed between conditions. In the morality condition, as before, beliefs about bad agents were more uncertain and volatile than beliefs about good agents. By contrast, in the competence condition beliefs about high and low skill agents were equally uncertain and volatile.

These findings suggest that asymmetries in belief volatility are specific to learning about *morally* bad agents.

Past research demonstrates that morally bad behaviors carry more weight than morally good behaviors during moral impression formation (Baumeister et al., 2001; Fiske, 1980; Mende-Siedlecki et al., 2013; Skowronski & Carlston, 1989). Because the bad agent (by definition) behaves worse than the good agent, it is unclear whether the learning asymmetry we observed is driven by asymmetric responses to the *behaviors* of the good and bad agents, or asymmetric responses to the *inferred character* of the agents. One possibility is that beliefs about bad agents *in general* are more uncertain and volatile, due to threat-evoked arousal. This possibility is consistent with work demonstrating that threatening cues enhance attention and accelerate information processing (Lojowska et al., 2019; Robinson et al., 2013). If interacting with a threatening agent destabilizes beliefs about them in general, beliefs about that bad agent's non-moral traits (e.g., competence) should also be more volatile than beliefs about a good agent's non-moral traits. Such a mechanism would be advantageous because it is useful to attend to and learn about all aspects of potentially harmful people, in order to build a richer model of those who pose a threat. We tested this hypothesis in a study where participants simultaneously inferred the morality and competence of a good and bad agent with similar levels of competence. Consistent with our predictions, beliefs about the bad agent's morality *and* competence than beliefs about these traits for the good agent. These findings suggest that our observation of more uncertain and volatile beliefs about the bad agent cannot be attributed to asymmetries in the choices that good and bad agents make, but rather seem to be a property of learning about bad agents in general.

One important implication of our model is that impressions of bad agents should be easily amenable in light of new information. Such a mechanism would make it possible to rapidly update an initially bad impression of someone if their behavior improves. We tested this in a study where

participants were randomized to learn about an agent who was initially either bad or good, but then began to make choices that were consistently either more or less moral than previously. Because beliefs about bad agents are more volatile, we predicted that participants would more strongly update their impressions of bad agents than good agents. We tested our hypothesis by comparing, for bad vs. good agents, the extent to which participants updated their impressions, defined as the difference between character ratings before vs. after the agents' preferences shifted.

Impression updating was stronger when the agent's behavior worsened than when it improved, consistent with past work on the negativity bias (Fiske, 1980; Pratto & John, 1991; Skowronski & Carlston, 1989). In addition and consistent with our predictions, impression updating was stronger for bad agents than good agents. The asymmetry in updating was particularly pronounced when behavior improved. Thus, the volatility of bad moral impressions may facilitate forgiveness by enabling initially bad impressions to be rapidly updated if behavior improves. This may reflect an adaptive mechanism for sustaining relationships when others sometimes behave badly. In other words, holding negative moral beliefs with some degree of uncertainty may be an important aspect of healthy social functioning.

Subsequent work has replicated the above findings in other contexts. For example, Bellucci and Park (2020) study a setting where participants repeatedly decide whether to take advice from agents who are either trustworthy (giving consistently honest advice) or untrustworthy (giving consistently dishonest advice). Crucially, halfway through the task, the trustworthy advisor begins behaving dishonestly, while the untrustworthy advisor begins behaving honestly. Participants changed their advice-taking behavior asymmetrically: they rapidly updated their beliefs about the initially untrustworthy agent and began following their advice. However, beliefs about the initially trustworthy agent remained stable; participants continued trusting this agent even when that advice was no longer valid (Bellucci & Park, 2020). Similarly, Lamba et al. (2020) study a repeated trust game setting where participants decide whether to entrust money with different agents whose trustworthiness

(operationalized as the amount of entrusted money returned to the participant) drifts slowly over the course of the experiment. One agent starts out trustworthy and gradually becomes untrustworthy, while another agent starts out untrustworthy and gradually becomes trustworthy. Paralleling the results of Bellucci and Park (2020), they find that initial trustworthiness impressions biased subsequent learning. Participants were slower to update their behavior toward initially trustworthy than initially untrustworthy agents (Lamba et al., 2020). Overall, these findings suggest that across a variety of settings, beliefs about initially untrustworthy agents are more uncertain and volatile than beliefs about initially trustworthy agents. However, it is important to note that all studies on this phenomenon to date have studied only moderately "good" and "bad" behaviors; future work is required to determine whether these findings will generalize to learning about more extreme behaviors. It also remains unclear what algorithms people actually use to form and update moral impressions, and how those algorithms are implemented at the neural level. One intriguing question for future work is whether changing the computational goal changes algorithms and implementations (Lockwood et al., 2020). If this is the case, then moral inference – which seems to involve relational goals that do not operate during non-social inference – may rely on unique brain systems.

3.4. Moral inference and exposure to community violence

Any discussion about the mechanisms for adaptive social functioning would be incomplete without considering the potential consequences that ensue when these mechanisms break down. Recently we have investigated moral inference in individuals exposed to violence. Over 75% of youth in the United States have faced some form of community violence over their lifetime (Finkelhor et al., 2013, 2015). Exposure to violence is a risk factor for a number of negative outcomes, including mental health problems (D. Baskin & Sommers, 2015; Fowler et al., 2009; Moffitt & Tank, 2013), interpersonal problems (Guo et al., 2013), antisocial and aggressive behavior (D. Baskin & Sommers, 2015; DuRant et

al., 1994; Fowler et al., 2009; Javdani et al., 2014), and involvement in the justice and social service systems (Hawkins et al., 2000).

Prominent theories about the relationship between exposure to violence and antisocial behavior stress the central role of learning (Bandura, 1978; Guerra et al., 2003; Huesmann & Kirwil, 2007; Ng-Mak et al., 2002, 2004). If learning about the morality of others is a crucial component of healthy social functioning, it may be disrupted by exposure to violence. Siegel et al. (2019) studied a sample of incarcerated men to investigate how exposure to violence affects the ability to learn about others' morality and use this information to adaptively modulate trust behavior.

The data suggest that exposure to violence adversely impacted some, but not all, components of moral inference. All participants were able to learn the agents' moral preferences and accurately predict their decisions, and developed more uncertain and volatile beliefs about bad agents relative to good agents, regardless of their past exposure to violence. However, only those with low levels of exposure to violence were able to use this information to make adaptive trust decisions, placing more trust in the good agent than the bad. Meanwhile, those with the highest exposure to violence trusted the good and bad agents equally. In particular, participants with high exposure to violence extended less trust than optimal when interacting with the good agent, which meant that they missed out on potential benefits they could have gained from that interaction.

The relationship between exposure to violence and maladaptive trusting behavior was mediated by disturbances in moral impression formation. Consistent with evidence that exposure to violence normalizes beliefs about harm (Ng-Mak et al., 2002), exposure to violence predicted more lenient impressions of the bad agent. Meanwhile, those with the highest exposure to violence made harsher evaluations of the good agent, consistent with evidence that individuals who have experienced violence themselves interpret the behavior of neutral actors as more hostile (Dodge et al., 1990). These

disturbances in impression formation and trust behavior, in turn, predicted participants' antisocial behaviors in prison – in particular, aggressive violations against other people.

Overall, these findings suggest that chronic exposure to violence leaves lasting effects on the ability to form subjective, global impressions of moral character that distinguish those we should avoid from those we should befriend. In turn, this may disrupt the ability to develop healthy social relationships with trustworthy individuals and increase the likelihood of trusting the "wrong" people. Despite being able to accurately learn and predict the helpful and harmful behaviors of others, individuals with high exposure to violence had difficulties translating those learned predictions into adaptive social decision-making.

3.5. Moral inference in Borderline Personality Disorder

Another population with disrupted social behavior is individuals with Borderline Personality Disorder (BPD), a serious mental illness characterized by marked disturbances in interpersonal relationships. In BPD, relationships are intense and unstable (American Psychiatric Association, 2013). Relative to healthy adults, BPD patients' social networks have a greater number of relationships that are terminated prematurely (Clifton et al., 2007), and patients often hold grudges and present difficulty forgiving others (Sansone et al., 2013; Thielmann et al., 2014).

Building and maintaining successful social relationships depends on the ability to form accurate beliefs about others' mental states (e.g., intentions, beliefs, desires) (Frith & Frith, 2012). However, research indicates that individuals with BPD are negatively biased and hypervigilant in their social inferences, often misinterpreting others' actions as threatening or hostile (Barnow et al., 2009; Fertuck et al., 2013, 2018; Nicol et al., 2013; Preißler et al., 2010; Unoka et al., 2011). Negatively biased inferences have adverse consequences on patients' social networks, leading to unstable relationships. For example, social interactions in BPD are associated with a pattern of rapid shifting from periods of

admiration to dislike of social partners (Bender & Skodol, 2007) and the termination of close relationships in response to even minor slights (Clifton et al., 2007).

These relational difficulties might be explained by a tendency for BPD patients to form overly rigid negative beliefs about the morality of others, in contrast to healthy adults who hold negative moral beliefs with uncertainty, allowing them to be flexibly updated in a way that facilitates forgiveness and relational stability. We tested this hypothesis in a study of moral inference in patients with BPD (Siegel et al., 2020). Using our moral inference task, we compared BPD patients with healthy adults matched for age, sex, and education. Both patients and healthy controls formed accurate beliefs about the morality of the good and bad agents. However, relative to healthy controls, BPD patients formed more certain beliefs about bad agents and were slower to update those beliefs. In addition, BPD patients formed more uncertain beliefs about the good agent and were faster to update those beliefs relative to the control group. Overall, the profile of moral inference in BPD patients differs markedly from healthy adults; patients have more confident and more rigid beliefs about putatively harmful social partners, but less confident and more flexible beliefs about putatively benevolent social partners.

Can psychiatric treatment ameliorate these maladaptive patterns of moral inference in BPD patients? Democratic Therapeutic Community (DTC) treatment is one of the most widespread psychosocial treatments in the UK with a strong focus on developing cooperative strategies to help patients effectively navigate their social environment (Whiteley, 2004). One of the strongest outcomes reported by participants following DTC is more pleasant social relations (Debaere et al., 2016). We investigated the impact of DTC on moral inference, comparing untreated BPD patients with BPD patients who had undergone DTC treatment. Consistent with the possibility that DTC treatment makes moral inference more adaptive, DTC-treated patients formed more uncertain, flexible beliefs about putatively harmful social partners than untreated patients (Siegel et al., 2020). In other words, moral inference in DTC-treated patients more closely resembled healthy adults than untreated patients. These findings

suggest that DTC may improve social interactions in BPD by increasing participants' openness to learning about partners who exhibited potentially threatening social interactions. Indeed, a key component of the DTC environment is that patients cannot "escape" or "write off" other members of the community.

3.6. Summary

Evolutionary models demonstrate that responding to wrongdoers with probabilistic forgiveness can facilitate the evolution of cooperation, but the cognitive mechanisms that enable the implementation of forgiving strategies are unknown. A series of studies showed that moral inference can be described by an asymmetric Bayesian updating mechanism, where beliefs about morally worse agents are more uncertain (and therefore more amenable to updating) than beliefs about morally better agents. Our model and data reveal a cognitive mechanism that enables flexible updating of beliefs about potentially threatening others, a mechanism that could facilitate forgiveness when initial bad impressions turn out to be inaccurate. This mechanism is disrupted in two populations characterized by interpersonal disturbances: individuals exposed to community violence and patients with BPD. The data reveal cognitive processes that may explain the emergence of maladaptive social behavior related to social trauma. Overall, these studies suggest that moral inference is optimized for social relationships and demonstrate the potential for combining behavioral paradigms with computational modelling as a tool for understanding both adaptive and maladaptive social functioning.

4. Moral inference from moral principles

Section 3 investigated how people make inferences about moral character from observing decisions to inflict pain on others for money. Such behaviors are typically seen as immoral because ethical systems universally prohibit harming others for personal gain (Gert, 2004). In this section, we explore how people infer moral character from judgments of how to resolve *moral dilemmas*: cases where ethical

systems disagree about what is the morally right course of action. What should one do, for example, if the only way to prevent a suspected major terrorist attack threatening thousands of lives is to torture the child of the suspected terrorist until she releases the information of where her father is?

This dilemma and others like it raise conflicts between two main families of ethical theories that have dominated philosophy and moral psychology: consequentialism and deontology. Consequentialist theories - of which utilitarianism is the most well-known exemplar (Bentham, 1983; Mill, 1863; Singer, 1993) - posit that only consequences matter when making moral decisions. In contrast, deontological ethical theories (Fried, 1978; Kant, 1797/2002; Rawls, 1971; Scanlon, 1998; W.D. Ross, 1930) posit that the notions of rights, duties, and obligations matter much more than the consequences of a decision. More specifically, utilitarianism differs from deontological ethics in at least two ways: it permits harming innocent individuals to maximize aggregate utility (instrumental harm), and it treats the interests of all individuals as equally important (impartial beneficence) (Everett & Kahane, 2020; Kahane et al., 2018). With respect to the dilemma above, utilitarian theories conclude that the torture is permissible, even required, because it will in this context bring about better consequences by avoiding a terrorist attack. In contrast, deontological theories argue that even if harming an innocent to save the lives of others maximizes overall welfare ("the Good"), this doesn't mean it is morally correct ("the Right") – however attractive such an action might be in this specific context, it is unacceptable because it violates human rights and prohibitions against torture.

In this section, we review studies of the moral inferences people make about others who endorse either utilitarian or deontological solutions to moral dilemmas. Following the conclusions of Section 3, this work reveals that moral inferences from moral principles also follows a relational logic: people are more likely to trust others who endorse moral principles that are adaptive for their particular social roles. For close relationship partners, people tend to prefer deontologists, perhaps because

deontological ethics is sensitive to the demands and obligations of closer relationships. For leadership roles, however, the picture is more complicated.

4.1. Moral inference from instrumental harm

Sacrificial "trolley-style" dilemmas have been widely used in moral psychology to investigate the psychology of *instrumental harm*: a willingness to cause harm in order to bring about better consequences overall. In the terrorist example above, for example, is it acceptable to cause harm to the child of a terrorist in order to bring about information that could prevent a deadly terrorist attack? Much work has focused on the psychological processes underlying (e.g. Greene et al., 2001, 2008), and the individual differences associated with (e.g. Conway et al., 2018; Kahane et al., 2015), such as endorsement of instrumental harm. A wealth of behavioral and neurobiological evidence has shown that participants' intuitive and automatic judgments tend to support a deontologically-consistent rejection of instrumental harm, whereas pro-sacrificial judgments are often the result of slow, deliberative cognitive processes (Greene et al., 2001, 2008; Koenigs et al., 2007). Precisely *why*, though, would moral intuitions more often align with a deontological rejection of sacrifice, rather than the option that would maximize the utility of outcomes? One way of beginning to understand this question comes from looking at the social function of such judgments: how does endorsing versus rejecting instrumental harm signal ones' suitability as a social partner?

Everett et al. (2016) investigated individuals' perceptions of people who made either characteristically deontological or utilitarian judgments in the footbridge dilemma: an instrumental harm dilemma that typically evokes deontological intuitions in most respondents (Cushman, Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Participants were presented with information about two "agents" who were asked to judge whether it was morally appropriate to push a man off a footbridge to stop an oncoming train from hitting five others, thereby killing him but saving

the five. The utilitarian agent judged it morally appropriate to push the man ("it is better to save five lives than one"), while the deontological agent rejected the sacrifice ("killing people is just wrong, even if it has good consequences"). Across a series of studies, we measured participants' moral impressions and trust in each agent, using both self-report measures of perceived morality and trustworthiness, as well as a behavioral measures of trust where participants were invited to entrust money to each agent in a one-shot trust game (Berg et al., 1995).

Across all studies, a highly consistent pattern emerged: participants perceived those who gave deontological responses to a sacrificial moral dilemma as more trustworthy than those who gave utilitarian responses, evidence both in self-reports and in behavior in a trust game. These results have been subsequently replicated in numerous studies by independent research groups (M. Brown & Sacco, 2017; Rom et al., 2017; Sacco et al., 2017) and pre-registered replication projects (Everett et al., 2018) with a variety of instrumental harm dilemmas. Moreover, these results are not attributable to participants simply preferring those who made similar judgments to them, for results are robust to controlling for participants' own moral judgments and while participants who make deontological judgments (the majority) tend to strongly prefer a deontological agent, those participants who made a utilitarian judgment (the minority) typically show no preference between the two agents (Everett et al., 2016).

Why are deontologists trusted less than utilitarians? Everett et al. (2016, 2018) argue there are two key distinctions between deontological and consequentialist ethics like utilitarianism that seem be particularly important in explaining the preference for deontologists. First, following of deontological rules leads to inherently more predictable behavior than utilitarian cost-benefit analyses. Second, deontological judgments relating to duties, obligations, and aversion to harm typically indicate more socially valued beliefs about others than the utilitarian acceptance of any-and-all actions in pursuit of the greater good. When it comes to selecting social partners and determining whom to trust, this makes

a preference for deontologists socially rational. In practice, the two go together: the utilitarian rejection of any constraints on the maximization of welfare means that there is no place for rights, duties, and respect for individual persons, and this makes cooperative behavior difficult to predict. For example, if by stealing your new laptop and selling it on the black market we could make a lot of money that we could donate to charities in the developing world to save children's lives, a utilitarian might argue that this is what we *should* do – regardless of whether we have previously made (potentially implicit) commitments not to steal from you. But expected adherence to such implicit commitments is critical when selecting a social partner for the purposes of cooperative exchange (e.g. friend, spouse, colleague).

The existing evidence suggests that both predictability and perceived socially valued beliefs may explain the preference for those who endorse deontological morality – though the perception of socially valued beliefs about respect for others might be more important. In their Study 2, for example, Everett et al. (2016) consider that if deontological agents are preferred over utilitarian agents because they are perceived as more committed to social cooperation, these preferences should be reduced if utilitarian agents reported their judgments as being very difficult to make – thereby indicating some level of commitment to cooperation. Someone who reports that it is easy to make a characteristically utilitarian, or pro-sacrificial, judgment can be interpreted as being high in utilitarianism (because they endorsed the sacrifice) but low in deontology (because there was little decision conflict with competing deontological motives). In contrast, someone who reports it is difficult to make the pro-sacrificial judgment can be interpreted as being high in both utilitarianism (because they endorsed the sacrifice) *and* deontology (because there was decision conflict with simultaneous deontological intuitions to not endorse the sacrifice; Conway & Gawronski, 2013). Therefore, to the extent that it is the presence of deontological intuitions that are perceived to indicate an emotional commitment to cooperation that is crucial for inferring trustworthiness, the preference for a deontologist over a utilitarian agent should be lessened

when the utilitarian agent reports difficulty in making the judgment. In contrast, if the preference for deontologists emerges purely from their predictability, then expressing difficulty should have little effect.

To test this, Everett et al (2016) asked participants to play a trust game with an agent who gave a deontological or utilitarian response to a sacrificial moral dilemma, but in this study added information that the target reported that their judgment was either "very difficult" or "very easy" to make. Their results showed that while deontologists were seen as more trustworthy overall, there was a significant interaction with choice difficulty. Breaking the interaction down, reported decision difficulty had no significant effect on perceived trust of deontological agents, and while not statistically significant, there was the suggestion that a utilitarian agent who said the decision was difficult was rated as more trustworthy than when they did not express such conflict.

Further evidence that the preference for deontologists over utilitarians is driven more by the perception of socially valued beliefs and responses to others than predictability *per se* comes from a comparison of trust in individuals who made different types of deontological (and utilitarian) judgments. Discussions of the trolley dilemma in moral psychology tend to focus on the deontological theory of Kant, who held that moral law consists of a set of maxims, or rules, that are categorical in nature, and that we are bound by duty to act in accordance with these categorical imperatives (Kant, 1797/2002). (In)famously, according to certain interpretations of Kant, we must always follow moral rules even when they will demonstrably cause harm – for example, if one was in WWII era Germany and was asked by the Nazis if they were hiding Jews in their attic. To lie would save the hidden Jews from a tragic death, but – at least in popular theory – Kant would require one to tell the truth, regardless of the consequences. But just as people do not trust utilitarians, it seems also unlikely that people would trust such a Kantian – someone who is so wedded to moral norms that they are willing to cause grievous harm to avoid breaking them. It is this (simplified) Kantian view - whereby certain acts are intrinsically

morally right or wrong - that predominates in the moral psychology literature when deontology is discussed. But this simplified account often ignores the critical roles of justice, duties, obligations, and rights that are central features of (neo-)Kantian ethics – and arguably the more important ones when it comes to understanding moral inference. Deontology is a broad church; might be there be particular kinds of deontological judgments that particularly signal trust?

We might expect deontological judgments rooted in the idea of social contracts to be particularly important for signaling trust, given the importance of cooperation in our evolutionary history. Some deontological thinkers have extended Kantian thinking: for example, by focusing on the idea of social contracts and the ways our actions can be justified to one another (Gauthier, 1986; Hobbes, 1668/1994; Parfit, 1984; Rawls, 1971; Scanlon, 1998). Of particular interest is recent theoretical work that has argued for the evolution of a *contractualist morality* by partner choice mechanisms (Baumard & Sheskin, 2015). Moral contractualism is a non-consequentialist ethical theory developed by Scanlon (1998), according to which moral actions are those that would result if we were to make fair and binding agreements – i.e., social contracts - from a point of view that respects our equal moral importance as rational autonomous agents. To the extent that people infer trustworthiness from sacrificial moral decisions because they indicate socially valued responses, this should be particularly so when someone signals moral intuitions that are consistent with the demands of justice and our mutual obligations to one another ("contractualist") as opposed to those that draw solely on specific acts being wrong regardless of the context ("categorical").

Everett et al. (2016) explored this in Studies 5a and 5b with a new dilemma that teased apart the following of rules with deontological notions of respect for personal autonomy – a so-called "soldier's dilemma." In this dilemma, a soldier is badly injured and caught in a trap, with the enemy fast approaching. The soldier cannot escape, and begs the troop leader not to leave him behind, as he will be cruelly tortured to death. Should the troop leader stab the soldier in the heart to prevent his suffering at

the hands of the enemy? Here, in contrast to the footbridge and sacrificial dilemmas used before, it is the *endorsement* – rather than rejection - of the sacrificial action that is consistent with a contractualist deontological ethical analysis: to sacrifice the person adheres to their wishes and requests, respecting their autonomy and wishes as a person. To tease apart how much moral inference of trust from sacrificial decisions is dependent on following predictable rules versus perceived socially valued responses and beliefs, Everett et al. presented participants with the Soldier's dilemma and manipulated both the decision and justification the agent gave, and whether the soldier themselves wanted to be sacrificed or not. One agent focused on relieving suffering, regardless of whether consent was given or not (consequentialist: "It is acceptable to kill someone if it reduces overall suffering"); one agent focused on killing being wrong, regardless of whether consent was given or not (categorical: "Killing people is just wrong, even if it has good consequences"); and two agents focused explicitly on autonomy and respecting the soldier's wishes (contractualist: ("It's right [wrong] to kill the soldier if that's [not] what they want, and it's the commander's duty to respect that"). Importantly, results showed that regardless of whether the decision made was to sacrifice or not, participants preferred the agent who endorsed the action that conformed to the soldier's wishes over those who endorsed the action that did not. Taken together, these findings highlight the way that perceptions of trustworthiness are sensitive not just to whether someone is moral, but the specific *kinds* of moral judgments and principles they endorse.

4.2. Moral inference from impartial beneficence

Much work in moral psychology investigating utilitarianism – the most widely discussed consequentialist theory - has utilized sacrificial trolley-style dilemmas, focusing on the psychological processes underlying and individual differences associated with instrumental harm. Yet there are other – more prototypical - ways in which utilitarianism, as an ethical theory, departs from common-sense moral intuitions. While utilitarianism permits instrumentally harming innocent individuals when this

maximizes aggregate utility, it also requires us to follow a principle of *impartial beneficence,* considering the welfare of all sentient beings equally (Everett & Kahane, 2020; Kahane et al., 2018). The utilitarian principle of impartial beneficence goes beyond "ordinary" morality not only with respect to how *much* we should sacrifice but also for *whose* sake: it tells us to impartially maximize the well- being of all sentient beings on the planet, not privileging compatriots, family members, or ourselves over strangers. The prominent utilitarian philosopher Peter Singer may have defended infanticide in some contexts (an example of instrumental harm), but his primary moral aims are the ones relating to impartial beneficence – for example, making great sacrifices to prevent the suffering of animals or people living in poverty around the world. According to the two-dimensional model of utilitarian psychology (Everett & Kahane, 2020; Kahane et al., 2018), instrumental harm and impartial beneficence are not just distinct conceptually but also psychologically, exhibiting distinct patterns of individual differences (Kahane et al., 2015, 2018) and appearing to rely on distinct psychological processes (Capraro et al., 2020).

Recent work has begun to shed light on how people infer moral character from impartial beneficence. People often approve of impartiality, preferring fairness to unfairness (Shaw, 2013; Tyler, 2000) but also judge others to have a better moral character when they act partially (helping a family member) rather than impartially (helping a stranger) (Hughes, 2017; McManus et al., 2020). What happens when partiality and beneficence conflict? How do we infer moral character from decisions about whether to provide a smaller benefit to a close friend or family member, or to provide a much larger benefit to strangers? Just as utilitarianism endorses instrumentally harming others to achieve the greater good, so too does it reject the idea of "special obligations", or the idea that we should help certain people by virtue of their close relationship with us. This means that just as with instrumental harm, in the domain of impartial beneficence there is a deep conflict between what we seek in a close social partner and what utilitarianism requires. Just as people distrust those who endorse utilitarian instrumental harm because they do not want social partners who will harm them in order to maximize

the greater good, might they also distrust others who endorse utilitarian impartial beneficence because they want social partners who will help them preferentially, even when it does not maximize the greater good?

Everett et al. (2018) investigated this across four studies in which we measured moral inferences about people who endorsed either utilitarian or deontological solutions to dilemmas involving either instrumental harm or impartial beneficence. For instrumental harm, we directly replicated the methods in Everett et al. (2016; see section 4.1 above). For impartial beneficence, we adapted some of the "greater good dilemmas" in Kahane and Everett et al. (2015). Should, for example, someone spend their weekend cheering up their lonely mother or instead help re-build houses for families who have lost their homes in a flood? Should a grandmother who has recently won a cash prize use the money to help her grandson repair his car, or instead donate to a charity that saves lives in the developing world? Using the same experimental design as Everett et al. (2016), we instructed participants to (ostensibly) interact with other participants who had read these dilemmas and made either a characteristically deontological decision that honored special obligations (i.e. spend time with the mother and give money to the grandson) or a characteristically utilitarian decision that focused on impartially maximizing overall outcomes (i.e. spend time re-building the stranger's damaged homes, or give the money to charity) (see **Table 1**).

Table 1. Reported judgments and justifications of the agent in Everett et al. (2018)

		Deontological Agent	Utilitarian Agent
Instrumental Harm	Footbridge	"I think that Amy should not push the large man to save the five workers. I know that by doing this she could stop the trolley and save more lives, but I think that killing people is just wrong even if it has good consequences."	"I think that Amy should push the large man to save the five workers. By doing this she could stop the trolley and save more lives, and I think that it is better to save many lives than just one"

	Vaccine	"I think that Amy should not inject the lab assistants with the substances to find out which is the vaccine. I know that by doing this she could find out which is the vaccine and therefore be able to save many more lives, but I think that killing people is just wrong, even if it has good consequences."	"I think that Amy should inject the lab assistants with the substances to find out which is the vaccine. By doing this she could find out which is the vaccine and therefore be able to save many more lives, and I think that it is better to save many lives than just one"
Impartial Beneficence	Spending Money	"I think that Amy should give the $500 to her grandson so that he could get his car fixed. I know that by donating the money to charity Amy would have the chance to bring about more happiness for more people, but I also think that respecting the duties she has to her grandson is more important."	"I think that Amy should give the $500 to the charity providing polio vaccinations in the developing world. I know that by donating the money to charity Amy would have the chance to bring about more happiness for more people, and I think this is more important than any duties she has to her grandson"
	Spending Time	"I think that Amy should spend the time with her mother instead of volunteering to build houses. I know that that by volunteering Amy would have the chance to bring about more happiness for more people, but I also think that respecting the duties she has to her mother is more important."	"I think that Amy should spend the time volunteering to build houses instead of spending the time with her mother. I know that by volunteering Amy would have the chance to bring about more happiness for more people, and I think this is more important than any duties she has to her mother"

As well as theoretically extending moral inference by considering impartial beneficence in addition to instrumental harm, we took the opportunity to investigate moral inference across a much greater range of dependent measures than has previously been used. We tested partner preference and trust by looking at two different economic games (the Trust Game and the Prisoner's Dilemma); several distinct dimensions along which the agent's character could be perceived (warmth; competence; morality); the different social roles in which the agent would be preferred (as a friend, a spouse, a boss, and as a political leader); and the different processes or motivations perceived to influence the agent's moral decision (reason vs. emotion; strategic considerations; altruistic motivations).

Across the studies, participants showed a consistent pattern of moral inference from instrumental harm, replicating the findings of Everett et al. (2016): utilitarians were seen as less moral and trustworthy, and preferred less as social partners, than deontologists. These results were observed across a variety of social roles. However, on the dimension of impartial beneficence, the results were more nuanced. While participants perceived those who endorsed impartial beneficence as consistently less loyal, there was no overall effect on perceived morality or trustworthiness – suggesting that the documented preference for people who show partiality towards kin was cancelled out by their simultaneous maximization of the greater good. In addition, moral inference from impartial beneficence was highly dependent on the relational role in question. We discuss these findings in more detail next.

4.3. Moral inference from moral principles is relationship-specific

There are growing reasons to think that the way we infer moral character from moral principles is highly relationship-specific, again highlighting the way that moral inference is flexibly used to facilitate the maintenance of social relationships. According to the theory of morality as relationship regulation (Rai & Fiske, 2011), there are different social relationships that people construct across cultures, with distinct moral obligations and prohibitions that these relationships entail. What might be moral in one relational context (e.g. someone charging a neighbor for babysitting their children) would be seen as immoral in another (e.g. a mother charging her son for help babysitting the grandchildren) (Clark et al., 2020; Earp et al., 2020). This suggests that moral inference from moral principles will be sensitive to the kind of social role or relationship that someone is judging for.

In the domain of instrumental harm, there is evidence that moral inference from sacrificial decisions is indeed sensitive to social roles. For example, those who endorse a utilitarian sacrifice are seen as less warm but more competent (Rom et al., 2017), and are preferred for organizational leadership positions like a hospital manager (Rom et al., 2017). Similarly, other work shows that people strategically choose

to endorse instrumental harm when the context favors competence-related traits, but are less likely to endorse instrumental harm when the context favors warmth-related traits (Rom & Conway, 2018). While such findings conflict with other results showing no such context sensitivity in what roles those who endorse instrumental harm are seen as suitable for (Everett et al., 2018), this remains a fruitful area for future research.

In the context of impartial beneficence, there is similarly evidence that moral inference is relationship-specific. For example, Everett et al. (2018) show that those who endorsed impartial beneficence are consistently viewed as being worse friends or spouses, but were not perceived as being worse political leaders (in fact, in two of the four studies the impartial utilitarian was expected to make a *better* political leader). While it makes sense for non-consequentialists to be favored for direct, interpersonal relationships (e.g. friend or spouse), it does seem much more reasonable to favor a consequentialist for distant, impersonal roles like a political leader, for whom their role can plausibly be described as requiring acting impartially to not favor one's own self-interest or the interests of those close to them. Consistent with this, recent work has shown that people do not endorse efficient, utilitarian maximization in charitable giving unless one is in a position of responsibility, like a political leader (Berman et al., 2018).

In recent work we have been investigating how endorsement of impartial beneficence and instrumental harm shape perceptions of political leaders in the COVID-19 pandemic (Everett, Colombatto et al., 2021). Previous work has shown that trust in leaders is a strong predictor of citizen compliance with a variety of public health policies (Levi & Stoker, 2000) – and during pandemics, trust in experts issuing public health guidelines is therefore a key predictor of compliance with those guidelines (Blair et al., 2017; Udow-Phillips & Lantz, 2020). Based on the work summarised thus far, one possible determinant of trust in leaders during a crisis is how they resolve moral dilemmas relating to instrumental harm and impartial beneficence. The COVID-19 pandemic has raised particularly stark

dilemmas of this kind. For example, a real instrumental harm dilemma that has played out over 2020 is whether to prioritize young and otherwise healthy people over the elderly and people with chronic illnesses when allocating scarce medical treatments. In the domain of impartial beneficence, we see dilemmas being played out in real time as leaders debate whether to prioritize their own citizens over people in other countries when allocating scarce resources. How might responses to these real-world moral dilemmas shape moral inference?

We tested the hypothesis that utilitarian responses to COVID-related moral dilemmas may erode or enhance trust in leaders relative to non-utilitarian approaches, depending on whether they concern instrumental harm or impartial beneficence. Specifically, we predicted that endorsement of instrumental harm would decrease trust in leaders, while endorsement of impartial beneficence would increase trust in leaders. Initial results from both the UK and USA collected in July 2020 provided support for our hypotheses. Leaders who made characteristically utilitarian decisions about instrumental harm (e.g. prioritizing the allocation of ventilators to younger and healthier people) were seen as less trustworthy. In contrast, leaders who made characteristically utilitarian decisions about impartial beneficence (e.g. endorsing sending medicine and personal protective equipment wherever it is required to do the most good, and not keeping it in one's own country) were seen as *more* trustworthy, consistent with the earlier findings by Everett et al. (2018). In November – December 2020, we replicated these findings in 22 countries across 6 continents (Australia, Brazil, Canada, Chile, China, Denmark, France, Germany, India, Israel, Italy, the Kingdom of Saudi Arabia, Mexico, the Netherlands, Norway, Singapore, South Africa, South Korea, Spain, United Arab Emirates, UK, and USA). Thus, in the context of real-world moral dilemmas, instrumental harm reduces trust in leaders, while impartial beneficence increases trust in leaders around the globe (Everett, Colombatto et al., 2021).

4.4. Summary

In summary, the work reviewed in this section shows that moral inference is sensitive not just to *whether* people behave morally, but also *what kind* of moral principles they endorse. Characteristically utilitarian judgments in the domain of instrumental harm erode trust: those who endorse harming one person to save many others are seen as less moral and trustworthy, and preferred less as social partners, across a variety of social roles. Meanwhile, endorsement of impartial beneficence can erode or enhance trust, depending on the relational context. For close social relationships characterized by duties and obligations, impartial beneficence is less desirable. However, for leadership roles, which require making decision that affect many, people find impartial beneficence to be more trustworthy.

5. Moral inference from 'non-moral' information

Thus far, we have reviewed how people infer moral character from information that is obviously morally relevant: decisions about harming others for personal gain, and decisions in moral dilemmas that pit different moral principles against one another. In this section, we consider the possibility that people make moral inferences from information that may not appear morally relevant on the surface. In daily life, we may frequently interact with people who, at least to our knowledge, have not committed severe or even minor moral transgressions. Yet, we may nevertheless make inferences about moral character even in the absence of overtly moral information. To the extent that 'non-moral' decisions about time, effort, uncertainty, or information seeking are correlated with moral preferences, it might be adaptive to infer moral character on the basis of such decisions.

5.1. Moral inference from decisions about time

Time preferences describe how people choose between smaller, sooner rewards and larger, delayed rewards. It is well established that the value of rewards are "discounted" over time, a phenomenon known as temporal (or delay) discounting. Those who prefer sooner rewards, at the expense of larger

rewards later, are called “steep discounters” (vs. “shallow discounters” who are said to exhibit patience). Are time preferences used as a cue to predict trustworthiness in cooperative encounters?

A relationship between time preferences and trustworthiness would not be surprising; many theorists have pointed out that people must incur short run costs in order to maximize benefits from long-term reciprocal relationships (Ainslie, 1992; R. Axelrod & Hamilton, 1981; Dewitte & De Cremer, 2001; Elster, 1985; Rachlin, 2002; Stevens et al., 2005). In fact, formal models of iterated cooperation even include a discount factor and predict that shallow discounters gain more utility from cooperating than steep discounters (R. Axelrod & Hamilton, 1981), since cooperating in the present round can help sustain cooperation across repeated future encounters (Green et al., 1995).

There is also evidence that shallow discounters are actually more trustworthy: after all, people who maximize long-run (vs. short-run) gains in decision-making tasks actually make more reputation-protective decisions (A. Vonasch & Sjåstad, 2019) and cooperate more, both in economic games and natural settings (Curry et al., 2008; Fehr & Leibbrandt, 2011; Harris & Madden, 2002; Yi et al., 2005). Specifically, time preferences appear to be positively associated with cooperation in iterated prisoner’s dilemma games and a one-shot public goods game (though the samples in these studies were small, with Ns ranging from 30-96). Perhaps even more compellingly, temporal discount rates elicited in the lab correlate with real world cooperative behavior: in a study of fishers and shrimpers from a lake community in Brazil, those who exhibited patient preferences in the lab (i.e., preferring two pralines later to one praline now) were less likely to over-exploit the common public resource, on average using nets with larger holes to allow younger fish/shrimp to escape (Fehr & Leibbrandt, 2011).

Some research suggests that a similar cognitive mechanism may underlie temporal decisions and social decisions: after all, decisions for future selves more closely resemble decisions for others than decisions for the present self (Pronin et al., 2008), and intriguingly, both intertemporal and social decisions engage the temporoparietal junction, an area associated with perspective-taking (Soutschek et

al., 2016). Additionally, social discounting, or deciding how to allocate resources among people closer (vs. farther) to one's self, follows a similar pattern to temporal discounting (Jones & Rachlin, 2006), and the two processes have similar neural substrates (Hill et al., 2017) and developmental trajectories (Garon et al., 2011). In a sense, kindness to one's "future self" may provide some clue as to how one would treat strangers.

Prudent personal decisions may also signal a general capacity to act in line with one's broader goals instead of succumbing to the immediate moment. A preferred trait in relationships, "self-control" may be key to positive social qualities such as overriding selfish impulses (Ainsworth & Baumeister, 2013; Gino et al., 2011; Gottfredson & Hirschi, 1990; Hirschi, 2004; Kocher et al., 2017; Pratt & Cullen, 2000; Righetti & Finkenauer, 2011). Accordingly, adults judge people who overcome conflict to do the right thing as more moral than those who did not experience conflict to begin with (Kee, 1969; Starmans & Bloom, 2016). And people who overcome small temptations, like eating unhealthy food (versus going to the gym) or spending money on entertainment (versus saving the money for college) are seen as more virtuous and trustworthy (Berman & Small, 2018; Righetti & Finkenauer, 2011).

5.2. Moral inference from decisions about effort

Morally good behavior often requires effort: helping a friend move, picking up a friend from the airport, and cleaning up a community park all involve the choice to spend effort on other people. Crucially, the level of effort that people spend on others is correlated with the effort they exert in pursuit of personal gains (Lockwood et al., 2017). So, the effort that people spend on themselves may be a cue to their prosocial capacities. How do people infer moral character from effortful choice, such as the choice between a high-effort, high-reward option versus a low-effort, low-reward option?

Especially in the U.S. and other cultures influenced by Protestantism, people seem to prefer hardworking others (Amos et al., 2019). Such a preference has been tied to cultural values, in particular

the Protestant Work Ethic (PWE; Weber, 2002): the valuation and moralization of work predominant in modern capitalism. German sociologist Max Weber (1864-1920) tied the emerging preference of work above other domains (e.g., leisure, time with family) to Lutheran beliefs on how practicing self-control and asceticism and accumulating wealth was "evidence" of being one of God's chosen ones (in contrast to earlier Calvinist teachings where indeed some people were chosen and some were not, but there was no visible evidence one way or another observable in one's lifetime). Social psychologists have since explored American and British people's individual differences in PWE (Mirels & Garrett, 1971) and associated attitudes towards work (Furnham, 1982).

In a series of experiments, Amos and colleagues explored how American participants infer moral character from effort-related traits (e.g., "indolent" vs. "hardworking"). Participants perceived "indolent" workers as less honest and more likely to cheat than "hard" workers (Amos et al., 2019). Even when hard work is unnecessary, people seem to infer moral character from decisions to exert effort. In studies of American, French, and South Korean participants, Celniker and colleagues probed moral evaluations of a character who either exerts high or low effort under conditions where effort does not generate any additional benefit. Even when effort is "unproductive," people who exerted effort were seen as more morally admirable (Celniker et al., 2020).

There are important limitations on the putative moralization of effort: such values may be culturally specific and also context specific -- after all, a "hardworking terrorist" is not seen as more moral than her lazy counterpart (Piazza et al., 2014). Additionally, if effort is valued, at least in some cultures, it is unclear how reputational concerns may influence effort-related decisions. If effort is interpreted as a cue to trustworthiness, do people exert effort in order to maintain appearances? After all, some extreme behaviors (e.g., mountain climbing, marathon training) are pursued for their high effort costs (Inzlicht et al., 2018) and donations to charity increase when coupled with an effortful activity, such as a 5k (Olivola & Shafir, 2013). Because most of the work on the moralization of effort has been conducted

with American participants, and psychological science has the proclivity to be biased toward some cultures (Henrich et al., 2010), future cross-cultural work can continue to examine moral character inferences from effortful decisions and its relation to public effort decisions.

5.3. Moral inference from decisions about ambiguity and risk

We are rarely certain of the outcomes of our actions. In the decision-making literature, uncertainty can take two forms: risk, where the underlying probability distribution of outcomes is known, and ambiguity, where the distribution is unknown. Individual preferences for risk and ambiguity are behaviorally and neurally dissociable (Hsu et al., 2005; Levy et al., 2009); people are generally both risk averse (Holt & Laury, 2002) and ambiguity averse (Ellsberg, 1961). Do attitudes toward uncertainty predict moral decisions? And if so, will certain kinds of decision makers (e.g., ambiguity tolerant, risk averse) be more trusted and preferred as social partners?

Vives and FeldmanHall (2018) note that many cooperative encounters, such as deciding whether to trust a stranger, are characterized more by ambiguity than risk. They measure participants' risk attitudes and ambiguity attitudes, and find that attitudes about ambiguity but not risk positively predict both decisions to cooperate in an iterated prisoner's dilemma and decisions to trust others (Vives & FeldmanHall, 2018). This suggests that in the absence of moral information, people may prefer social partners who have a higher tolerance for ambiguity. Supporting this idea, people feel more gratitude, infer more kindness, and are more likely to trust those who are willing to help when the cost of helping is ambiguous (Jordan, Hoffman, Nowak, et al., 2016; Xiong et al., 2020). A tolerance for ambiguity may also be related to holding more uncertain and volatile beliefs about potentially harmful others (Siegel et al., 2018). Both ambiguity tolerance and harmful belief volatility may be important for the ability to build and maintain healthy social relationships.

Meanwhile, there is growing evidence for a link between risk preferences and prosocial behavior. Kameda and colleagues (2016) explored the possibility that decisions about resource allocation and risk are related: indeed, most participants who maximized the minimum payoff in resource allocation decisions (following a Rawlsian "maximin" strategy) also maximized the minimum outcome in risky decisions, and vice versa -- most who maximized overall welfare in resource allocation also chose the option with the highest expected value in the risk task. Concern for the worst off party and the worst possible outcome were correlated in the task and were both associated with activity in the temporoparietal junction (TPJ), an area associated with perspective-taking (Kameda et al., 2016). Similarly, Muller and Rau (2016) found that more risk-averse participants were more likely to give fair and equitable allocations ("inequality averse") in a modified version of the Dictator Game (Müller & Rau, 2016). These findings suggest that risk aversion might serve as a cue for inferring moral character.

5.4. Moral inference from epistemic behaviors

A growing area of research concerns social evaluations and inferences from epistemic behavior: do people infer moral traits from epistemic behavior such as question-asking and information search? A source who has reliable information or knowledge about how things work can be a valuable asset to others (Danovitch & Keil, 2007; Lutz & Keil, 2002) but also has the potential to be misleading (Koenig et al., 2019; Richard et al., 2005; Sperber et al., 2010). Since engaging in certain kinds of epistemic behaviors, such as explanation, can be valuable for learning and teaching (Lombrozo, 2006), including moral lessons (Walker & Lombrozo, 2017), such behaviors may be moralized and/or preferred in teachers or friends.

Indeed, there is evidence we pay attention to others' epistemic qualities: children as young as 3-4 years old start to exhibit "selective imitation" in using information about informants' past accuracy and character traits when deciding whom to trust (Birch et al., 2008; Clement et al., 2004) and ask for advice

(Lane et al., 2013), and children infer that accurate informants are more prosocial (Brosseau-Liard & Birch, 2010). Children and adults also infer character traits on the basis of question-asking ability, predicting that good question askers are friendlier (Simone, n.d.).

Adults also use information-seeking behavior, such as the pursuit and/or denial of further evidence and explanations, as a cue to moral traits. Gill and Lombrozo (2019) presented American adults with a story about a character who learned about a topic (on near-death experiences or the shroud of Turin) and, depending on condition, the character either decided to pursue or forgo further inquiry on the topic, in the form of additional evidence or explanation. The topics-- near death experiences and the Shroud of Turin -- were chosen so that they could be manipulated to be framed in either a religious or scientific manner. But regardless of framing, participants rated the character who pursues (vs. forgoes) inquiry, whether in the form of evidence or explanation, as more moral and trustworthy.

5.5. Summary

There is growing evidence that people infer moral character from behaviors that are not explicitly moral. The data so far suggest that people who are patient, hard-working, tolerant of ambiguity, risk-averse, and actively open-minded are seen as more moral and trustworthy. While at first blush this collection of preferences may seem arbitrary, considering moral inference from a relational perspective reveals a coherent logic. All of these preferences are correlated with cooperative behavior, and comprise traits that are desirable for long-term relationship partners. Reaping the benefits of long-term relationships requires patience and a tolerance for ambiguity: sometime people make mistakes despite good intentions. Erring on the side of caution and actively seeking evidence to inform decision-making in social situations not only helps prevent harmful outcomes (Kappes et al., 2019), but also signals respect: social life is fraught with uncertainty (FeldmanHall & Shenhav, 2019; Kappes et al., 2019), and assuming we know what's best for another person can have bad consequences, even when

our intentions are good. If evidence continues to suggest that certain types of non-moral preferences are preferred in social partners, partner choice mechanisms may explain the prevalence of those preferences in the broader population.

6. Conclusions and future directions

We choose who to trust or avoid based on inferences about whether they are likely to help or harm us. Successfully inferring the moral character of others is crucial for initiating and maintaining healthy social relationships. When moral inference breaks down, we might place our trust in the wrong people, or prematurely end relationships because we incorrectly infer our partner means us harm.

Social psychologists have long studied how we form impressions of trustworthiness and morality, establishing that information relevant to assessing moral character dominates other kinds of information in person perception, and documenting a negativity bias whereby bad behaviors carry more weight than good behaviors in impression formation. More recent work has leveraged computational methods to build on this knowledge, identifying an important asymmetry in forming impressions of putatively harmful versus helpful others. When another person behaves in harmful or exploitative ways, we hold beliefs about that person with more uncertainty, enabling us to update those beliefs more rapidly in case they turn out to be better – or worse – than we originally thought. Bad impressions are not indelible; instead, they are held loosely, enabling forgiveness for accidental or rare transgressions. Meanwhile, we form more stable beliefs about people who behave generously and reciprocate our trust. This asymmetry in moral inference may be a critical feature of resilient social relationships, enabling long-term cooperative partnerships to persevere even when people sometimes defect by mistake. Accordingly, asymmetric moral inference is disrupted in clinical populations with difficulties sustaining healthy social relationships.

When deciding whom to trust, we not only care about *whether* people are moral, but also *what kind* of morality they endorse. Deontological morality broadly prohibits harming others and acknowledges that some relationships confer special obligations to help, while utilitarian morality permits harming others "for the greater good" and demands that we help others impartially, without privileging anyone. We find that people prefer to trust deontologists over utilitarians, particularly for close relationships (like a spouse or a friend) relative to more distant roles such as political leaders. Moreover, growing evidence suggests that people infer trustworthiness from behaviors that are not obviously moral: decisions about time, effort, uncertainty, and information seeking. The types of non-moral decisions that signal trust are not random, but rather represent a collection of preferences that are desirable for long-term social relationships. Overall, then, this research reveals the relational logic of moral inference: we are especially sensitive to information that bears on whether someone will make a reliable partner, and we update our beliefs about others in a way that promotes cooperation.

Future research on moral inference can build on these insights in a number of new directions. One intriguing set of questions concerns moral *self*-inference: how do we come to know our *own* moral character? It is well established that beliefs about the self are biased. Most people believe they are better than average across a variety of characteristics (Sedikides et al., 2003; Zell et al., 2019), and morality is no exception: people generally view themselves as more generous, kind and altruistic than they actually are (Allison et al., 1989; Epley & Dunning, 2000). Some moral self-enhancement might arise from biased memory processes: there is evidence that memories about one's own unethical behaviors are less vivid and accurate then memories about good deeds (Carlson et al., 2020; Kouchaki & Gino, 2016; Saucet & Villeval, 2019). In addition, moral self-enhancement might arise naturally from the same kinds of asymmetric belief updating processes that characterize moral inference about others. If our own harmful behaviors cause our beliefs about ourselves to become more uncertain and volatile, then subsequent moral behaviors could have a disproportionate impact on our moral self-beliefs, rapidly

restoring our moral self-image. This process could help explain the phenomenon of *moral compensation,* where people are more motivated to behave morally following a moral transgression (Gneezy et al., 2012; Zhong et al., 2009). By contrast, performing a good deed could shift moral self-beliefs in a positive direction, where belief updating is slower. Positive prior moral self-beliefs could be more resistant to change following moral transgressions, perhaps explaining the phenomenon of *moral licensing,* where people are more likely to transgress following a good deed (Mullen & Monin, 2016). Finally, it is worth noting that a number of clinical disorders are characterized by negative, unstable beliefs about the self, and many empirically validated therapies are based on the notion that psychiatric disorders arise from aberrant (moral) self-representations (Moutoussis et al., 2014). A deeper understanding of how people form and update moral self-beliefs may therefore be an important component of developing more effective treatments for psychiatric disorders. More broadly, understanding moral self-inference might be key to developing interventions that facilitate moral learning: people are only motivated to change their behavior if they believe that they need to change.

A second set of questions concerns how moral inference operates in the modern digital age. Artificially intelligent (AI) machines are increasingly tasked with decisions that have profound moral implications. For example, AI is being used to allocate scarce medical resources, inform parole decisions, and guide autonomous vehicles. How do people infer the moral "character" of AI? And how do moral inferences about AI relate to trust in these new technologies? There is some evidence that moral inferences about AI follows similar principles to human moral inference, such as mistrust in utilitarians. For instance, in dilemmas where autonomous vehicles must decide whether to sacrifice their passengers to save a larger number of pedestrians (a utilitarian strategy), most people would prefer to buy and ride in non-utilitarian vehicles that protect their passengers at all costs (Bonnefon et al., 2016). One intriguing possibility is that moral inference is relationship-specific for AI, just as it is for humans. People might prefer non-utilitarian AI for machines they interact with closely, like a personal car, but prefer

utilitarian AI for applications that affect broader swathes of the population, like an algorithm informing government policies for disaster relief. In any case, public acceptance of AI will depend on a deeper understanding of these issues.

In addition, algorithms are increasingly mediating our interactions with one another. As of this writing, there are over 3 billion active users of online social networks like Facebook, Instagram and Twitter combined, many of whom use the platforms daily for news consumption and political engagement. Moralized content (e.g., expressions of outrage) is especially likely to be shared online (Brady et al., 2017, 2020), and newsfeed algorithms prioritize and amplify this content (Brady et al., 2021; M. J. Crockett, 2017). Notably, people infer moral character from outrage expressions (Jordan, Hoffman, Bloom, et al., 2016; Jordan & Rand, 2019). Because people are more likely to encounter outrageous content online than offline (Crockett, 2017), moral inference might be operating more frequently and more intensively in our online than offline interactions. This has important implications for how we perceive others: if online interactions provide more information relevant to moral inference than offline interactions, we might be faster to form moral impressions of people we interact with online than offline. But how accurate are these impressions? Our newsfeeds do not show us everything expressed by everyone in our networks; they show us a cherry-picked sample, displaying only those expressions that the algorithms expect to grab our attention. Consider a hypothetical Twitter user, Alex, who tweets frequently and expresses outrage only once out of every 100 posts. But the newsfeed algorithm ignores the 99% neutral posts and promotes just the outrage expressions. To Alex's followers, Alex is all outrage all the time. But this moral inference is based on incomplete data – not only are Alex's followers ignorant to 99% of Alex's posts online, they probably know even less about how Alex behaves offline. On aggregate, this process could result in widespread misperceptions of moral character.

This hypothesis is related to a third, foundational question about moral inference. We know that moral behavior is highly sensitive to situational factors (Aquino et al., 2009; Ellemers et al., 2019; Gino,

2015; Hofmann et al., 2014; Tsang, 2002). In other words, there is very little evidence for the existence of *true* "good guys" and "bad guys." And yet, we cannot help but essentialize morality, considering moral characteristics to be immutable, innate, and central to identity (Heiphetz, 2019; Heyman & Gelman, 2000; Newman et al., 2014; Strohminger & Nichols, 2014). We are constantly on the lookout for evidence to guide moral inference, and moral inference seems to operate as if stable moral dispositions exist and are discoverable via observation. Why are we built this way? Given the overwhelming amount of data we encounter in our daily lives that shows us just how variable social behavior is, why do we nevertheless continue to divide up the world into heroes and villains? What are the benefits of moral inference operating in this way, and what are the costs? Perhaps the benefits of binarizing moral inference outweighed its costs in the environment of ancestral humans. But this may no longer be the case today, where moral information can ricochet around the globe with lightning speed and new technologies nudge us to rapidly form strong moral impressions of others whom we may never meet face-to-face. By continuing to study the computational mechanisms and relational logic of moral inference, we can begin to answer these important questions.

References

Ainslie, G. (1992). *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. Cambridge University Press.

Ainsworth, S. E., & Baumeister, R. F. (2013). Cooperation and fairness depend on self-regulation. *Behavioral and Brain Sciences*, *36*(1), 79–80. https://doi.org/10.1017/S0140525X12000696

Aksoy, O., & Weesie, J. (2014). Hierarchical Bayesian analysis of outcome- and process-based social preferences and beliefs in Dictator Games and sequential Prisoner's Dilemmas. *Social Science Research*, *45*, 98–116. https://doi.org/10.1016/j.ssresearch.2013.12.014

Albert Bandura. (1978). Social Learning Theory of Aggression. *Journal of Communication*, *28*(3), 12–29. https://doi.org/10.1111/j.1460-2466.1978.tb01621.x

Alexander, R. (1987). *The biology of moral systems*. Aldine de Gruyter.

Allison, S. T., Messick, D. M., & Goethals, G. R. (1989). On Being Better but not Smarter than Others: The Muhammad Ali Effect. *Social Cognition*, *7*(3), 275–295. https://doi.org/10.1521/soco.1989.7.3.275

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*.

Amos, C., Zhang, L., & Read, D. (2019). Hardworking as a Heuristic for Moral Character: Why We Attribute Moral Values to Those Who Work Hard and Its Implications. *Journal of Business Ethics*, *158*(4), 1047–1062. https://doi.org/10.1007/s10551-017-3725-x

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394–400. https://doi.org/10.1037/h0022280

Aquino, K., Freeman, D., Reed, A., Felps, W., & Lim, V. K. G. (2009). Testing a social-cognitive model of moral behavior: The interactive influence of situations and moral identity centrality. *Journal of Personality and Social Psychology*, *97*(1), 123–141. https://doi.org/10.1037/a0015406

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Axelrod, R. M. (2006). *The Evolution of Cooperation*. Basic Books.

Barnow, S., Stopsack, M., Grabe, H. J., Meinke, C., Spitzer, C., Kronmüller, K., & Sieswerda, S. (2009). Interpersonal evaluation bias in borderline personality disorder. *Behaviour Research and Therapy*, *47*(5), 359–365. https://doi.org/10.1016/j.brat.2009.02.003

Baskin, D., & Sommers, I. (2015). Trajectories of Exposure to Community Violence and Mental Health Symptoms Among Serious Adolescent Offenders. *Criminal Justice and Behavior*, *42*(6), 587–609. https://doi.org/10.1177/0093854814556882

Baskin, T. W., & Enright, R. D. (2004). Intervention Studies on Forgiveness: A Meta-Analysis. *Journal of Counseling & Development*, *82*(1), 79–90. https://doi.org/10.1002/j.1556-6678.2004.tb00288.x

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(01), 59–78. https://doi.org/10.1017/S0140525X11002202

Baumard, N., & Sheskin, M. (2015). Partner choice and the evolution of a contractualist morality. In J. Decety & T. Wheatley (Eds.), *The Moral Brain: A Multidisciplinary Perspective* (pp. 35–46). MIT Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Bellucci, G., & Park, S. Q. (2020). Honesty biases trustworthiness impressions. *Journal of Experimental Psychology: General*, *149*(8), 1567–1586. https://doi.org/10.1037/xge0000730

Bender, D. S., & Skodol, A. E. (2007). Borderline Personality as a Self-Other Representational Disturbance. *Journal of Personality Disorders*, *21*(5), 500–517. https://doi.org/10.1521/pedi.2007.21.5.500

Bentham, J. (1983). *The collected works of Jeremy Bentham: Deontology, together with a table of the springs of action ; and the article on utilitarianism*. Oxford University Press.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Berman, J. Z., Barasch, A., Levine, E. E., & Small, D. A. (2018). Impediments to Effective Altruism: The Role of Subjective Preferences in Charitable Giving. *Psychological Science*, *29*(5), 834–844. https://doi.org/10.1177/0956797617747648

Berman, J. Z., & Small, D. A. (2018). Discipline and desire: On the relative importance of willpower and purity in signaling virtue. *Journal of Experimental Social Psychology*, *76*, 220–230. https://doi.org/10.1016/j.jesp.2018.02.007

Birch, S. A. J., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, *107*(3), 1018–1034. https://doi.org/10.1016/j.cognition.2007.12.008

Blair, R. A., Morse, B. S., & Tsai, L. L. (2017). Public health and public trust: Survey evidence from the Ebola Virus Disease epidemic in Liberia—ScienceDirect. *Social Science & Medicine*, *172*, 89–97.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Brady, W. J., Crockett, M., & Bavel, J. J. V. (2020). The MAD Model of Moral Contagion: The role of motivation, attention and design in the spread of moralized content online. *Perspectives on Psychological Science*. https://doi.org/10.31234/osf.io/pz9g6

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. (2021). *How social learning amplifies moral outrage expression in online social networks*. PsyArXiv. https://doi.org/10.31234/osf.io/gf7t5

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Bavel, J. J. V. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*(2), 135–143. https://doi.org/10.1002/ejsp.744

Brambilla, M., Sacchi, S., Pagliaro, S., & Ellemers, N. (2013). Morality and intergroup relations: Threats to safety and group image predict the desire to interact with outgroup and ingroup members. *Journal of Experimental Social Psychology*, *49*(5), 811–821. https://doi.org/10.1016/j.jesp.2013.04.005

Brañas-Garza, P., Rodríguez-Lara, I., & Sánchez, A. (2017). Humans expect generosity. *Scientific Reports*, *7*(1), 42446. https://doi.org/10.1038/srep42446

Briscoe, M. E., Woodyard, H. D., & Shaw, M. E. (1967). Personality impression change as a function of the favorableness of first impressions. *Journal of Personality*, *35*(2), 343–357. https://doi.org/10.1111/j.1467-6494.1967.tb01433.x

Brosseau-Liard, P. E., & Birch, S. A. J. (2010). 'I bet you know more and are nicer too!': What children infer from others' accuracy. *Developmental Science*, *13*(5), 772–778. https://doi.org/10.1111/j.1467-7687.2009.00932.x

Brown, M., & Sacco, D. F. (2017). Is pulling the lever sexy? Deontology as a downstream cue to long-term mate quality. *Journal of Social and Personal Relationships*, 0265407517749331. https://doi.org/10.1177/0265407517749331

Brown, R. P. (2003). Measuring Individual Differences in the Tendency to Forgive: Construct Validity and Links with Depression. *Personality and Social Psychology Bulletin*, *29*(6), 759–771. https://doi.org/10.1177/0146167203029006008

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond Bipolar Conceptualizations and Measures: The Case of Attitudes and Evaluative Space. *Personality and Social Psychology Review*, *1*(1), 3–25. https://doi.org/10.1207/s15327957pspr0101_2

Capraro, V., Everett, J. A. C., & Earp, B. D. (2020). Priming intuition decreases instrumental harm but not impartial beneficence. *Journal of Experimental Social Psychology*.

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, *11*(1), 2100. https://doi.org/10.1038/s41467-020-15602-4

Carr, C. T., & Walther, J. B. (2014). Increasing Attributional Certainty via Social Media: Learning About Others One Bit at a Time. *Journal of Computer-Mediated Communication*, *19*(4), 922–937. https://doi.org/10.1111/jcc4.12072

Celniker, J., Gregory, A., Koo, H., Piff, P. K., Ditto, P. H., & Shariff, A. (2020). *The Moralization of Unproductive Effort*. PsyArXiv. https://doi.org/10.31234/osf.io/nh9ax

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience*, *13*(6), 637–647. https://doi.org/10.1080/17470919.2018.1518834

Clark, M. S., Earp, B. D., & Crockett, M. J. (2020). Who are "we" and why are we cooperating? Insights from social psychology. *The Behavioral and Brain Sciences*, *43*, e66. https://doi.org/10.1017/S0140525X19002528

Clement, F., Koenig, M., & Harris, P. (2004). The Ontogenesis of Trust. *Mind and Language*, *19*(4), 360–379. https://doi.org/10.1111/j.0268-1064.2004.00263.x

Clifton, A., Pilkonis, P. A., & McCarty, C. (2007). Social Networks in Borderline Personality Disorder. *Journal of Personality Disorders*, *21*(4), 434–441. https://doi.org/10.1521/pedi.2007.21.4.434

Cone, J., & Ferguson, M. J. (2015). He Did What?: The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, *179*, 241–265. https://doi.org/10.1016/j.cognition.2018.04.018

Cornwell, B. R., Garrido, M. I., Overstreet, C., Pine, D. S., & Grillon, C. (2017). The Unpredictive Brain Under Threat: A Neurocomputational Account of Anxious Hypervigilance. *Biological Psychiatry*, *82*(6), 447–454. https://doi.org/10.1016/j.biopsych.2017.06.031

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology*, *92*(2), 208–231. https://doi.org/10.1037/0022-3514.92.2.208

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

Crockett, Molly J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320–17325. https://doi.org/10.1073/pnas.1408988111

Curry, O. S., Price, M. E., & Price, J. G. (2008). Patience is a virtue: Cooperative people have lower discount rates. *Personality and Individual Differences*, *44*(3), 780–785. https://doi.org/10.1016/j.paid.2007.09.023

Danovitch, J. H., & Keil, F. C. (2007). Choosing between hearts and minds: Children's understanding of moral advisors. *Cognitive Development*, *22*(1), 110–123. https://doi.org/10.1016/j.cogdev.2006.07.001

De Bruin, E. N. M., & van Lange, P. A. M. (2000). What People Look for in Others: Influences of the Perceiver and the Perceived on Information Selection. *Personality and Social Psychology Bulletin*, *26*(2), 206–219. https://doi.org/10.1177/0146167200264007

Debaere, V., Vanheule, S., Van Roy, K., Meganck, R., Inslegers, R., & Mol, M. (2016). Changing encounters with the other: A focus group study on the process of change in a therapeutic community. *Psychoanalytic Psychology*, *33*(3), 406–419. https://doi.org/10.1037/a0036862

Dewitte, S., & De Cremer, D. (2001). Self-control and cooperation: Different concepts, similar decisions? A question of the right perspective. *The Journal of Psychology*, *135*(2), 133–153. https://doi.org/10.1080/00223980109603686

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., & Stephan, K. E. (2014). Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLoS Comput Biol*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Dibbets, Pauline, Adolphs, Laura, Close, Ingeborg, Herings, Anke, Kiggen, Maiken, Kinneging, Maaike, Loffler, Leonie, Nijssen, Yonne, Schulte-Ostermann, Michel, van Schaaik, Patrick, Section Clinical Psychology, & RS: FPN CPS III. (2012). Reversal of Attitude: The Influence of Counter-Attitudinal Information. *Journal of Social Sciences*, *8*(3), 390–396.

Dodge, K. A., Bates, J. E., & Pettit, G. S. (1990). Mechanisms in the cycle of violence. *Science*, *250*(4988), 1678–1683. https://doi.org/10.1126/science.2270481

Dunbar, R. I. M. (2004). Gossip in Evolutionary Perspective. *Review of General Psychology*, *8*(2), 100–110. https://doi.org/10.1037/1089-2680.8.2.100

DuRant, R. H., Pendergrast, R. A., & Cadenhead, C. (1994). Exposure to violence and victimization and fighting behavior by urban black adolescents. *Journal of Adolescent Health*, *15*(4), 311–318. https://doi.org/10.1016/1054-139X(94)90604-1

Earp, B. D., McLoughlin, K., Monrad, J., Clark, M. S., & Crockett, M. (2020). *How social relationships shape moral judgment*. PsyArXiv. https://doi.org/10.31234/osf.io/e7cgq

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. https://doi.org/10.1038/nn.3428

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019). The Psychology of Morality: A Review and Analysis of Empirical Studies Published From 1940 Through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. https://doi.org/10.1177/1088868318811759

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669. https://doi.org/10.2307/1884324

Elster, J. (1985). Weakness of Will and the Free-Rider Problem. *Economics & Philosophy*, *1*(2), 231–265. https://doi.org/10.1017/S0266267100002480

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *Journal of Cognitive Neuroscience*, *19*(9), 1508–1519. https://doi.org/10.1162/jocn.2007.19.9.1508

Epley, N., & Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology*, *79*(6), 861–875. https://doi.org/10.1037//0022-3514.79.6.861

Everett, J. A. C., Colombatto, C., Awad, E., Boggio, P., Bos, B., Brady, W. J., Chawla, M., Chituc, V., Chung, D., Drupp, M., Goel, S., Grosskopf, B., Hjorth, F., Ji, A., Lin, Y., Ma, Y., Maréchal, M., Mancinelli, F., Mathys, C., … Crockett, M. J. (2021). Moral dilemmas and trust in leaders during a global health crisis. *Nature Human Behaviour*.

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004

Everett, J. A. C., & Kahane, G. (2020). Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.11.012

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. https://doi.org/10.1037/xge0000165

Fehr, E., & Leibbrandt, A. (2011). A field study on cooperativeness and impatience in the Tragedy of the Commons. *Journal of Public Economics*, *95*(9), 1144–1155. https://doi.org/10.1016/j.jpubeco.2011.05.013

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences*, *115*(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435. https://doi.org/10.1038/s41562-019-0590-x

Fertuck, E. A., Grinband, J., Mann, J. J., Hirsch, J., Ochsner, K., Pilkonis, P., Erbe, J., & Stanley, B. (2018). Trustworthiness appraisal deficits in borderline personality disorder are associated with prefrontal cortex, not amygdala, impairment. *NeuroImage: Clinical*, 101616. https://doi.org/10.1016/j.nicl.2018.101616

Fertuck, E. A., Grinband, J., & Stanley, B. (2013). Facial trust appraisal negatively biased in borderline personality disorder. *Psychiatry Research*, *207*(3), 195–202. https://doi.org/10.1016/j.psychres.2013.01.004

Finkelhor, D., Turner, H. A., Shattuck, A., & Hamby, S. L. (2013). Violence, Crime, and Abuse Exposure in a National Sample of Children and Youth: An Update. *JAMA Pediatrics*, *167*(7), 614–621. https://doi.org/10.1001/jamapediatrics.2013.42

Finkelhor, D., Turner, H. A., Shattuck, A., & Hamby, S. L. (2015). Prevalence of Childhood Exposure to Violence, Crime, and Abuse: Results From the National Survey of Children's Exposure to Violence. *JAMA Pediatrics*, *169*(8), 746–754. https://doi.org/10.1001/jamapediatrics.2015.0676

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. https://doi.org/10.1037/0022-3514.38.6.889

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fouragnan, E. (2013). *The Neural Computation of Trust and Reputation* [Phd, University of Trento]. http://eprints-phd.biblio.unitn.it/970/

Fowler, P. J., Tompsett, C. J., Braciszewski, J. M., Jacques-Tiura, A. J., & Baltes, B. B. (2009). Community violence: A meta-analysis on the effect of exposure and mental health outcomes of children and adolescents. *Development and Psychopathology*, *21*(1), 227–259. https://doi.org/10.1017/S0954579409000145

Freedman, J. L., & Steinbruner, J. D. (1964). Perceived choice and resistance to persuasion. *The Journal of Abnormal and Social Psychology*, *68*(6), 678–681.

Fried, C. (1978). *Right and Wrong*. Harvard University Press.

Frith, C. D., & Frith, U. (2012). *Mechanisms of Social Cognintion*. *63*, 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Fudenberg, D., Rand, D. G., & Dreber, A. (2012). Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *The American Economic Review*, *102*(2), 720–749. https://doi.org/10.1257/aer.102.2.720

Furnham, A. (1982). The Protestant work ethic and attitudes towards unemployment. *Journal of Occupational Psychology*, *55*(4), 277–285. https://doi.org/10.1111/j.2044-8325.1982.tb00101.x

Garon, N., Johnson, B., & Steeves, A. (2011). Sharing with others and delaying for the future in preschoolers. *Cognitive Development*, *26*(4), 383–396. https://doi.org/10.1016/j.cogdev.2011.09.007

Gartner, J. (1988). The capacity to forgive: An object relations perspective. *Journal of Religion and Health*, *27*(4), 313–320. https://doi.org/10.1007/BF01533199

Gauthier, D. P. (1986). *Morals by Agreement*. Oxford University Press.

Gert, B. (2004). *Common Morality: Deciding What to Do*. Oxford University Press.

Gino, F. (2015). Understanding ordinary unethical behavior: Why people who value morality act immorally. *Current Opinion in Behavioral Sciences*, *3*, 107–111. https://doi.org/10.1016/j.cobeha.2015.03.001

Gino, F., Schweitzer, M. E., Mead, N. L., & Ariely, D. (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 191–203. https://doi.org/10.1016/j.obhdp.2011.03.001

Gneezy, U., Imas, A., & Madarasz, K. (2012). *Conscience Accounting: Emotional Dynamics and Social Behavior*. http://works.bepress.com/kristof_madarasz/22

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime* (pp. xvi, 297). Stanford University Press.

Green, L., Price, P. C., & Hamburger, M. E. (1995). Prisoner's Dilemma and the Pigeon: Control by Immediate Consequences. *Journal of the Experimental Analysis of Behavior*, *64*(1), 1–17. https://doi.org/10.1901/jeab.1995.64-1

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Griffiths, T. L., Kemp, C., & B. Tenenbaum, J. (2008). *Bayesian models of cognition*. https://doi.org/10.1184/R1/6613682.v1

Grim, P. (1995). The greater generosity of the spatialized prisoner's dilemma. *Journal of Theoretical Biology*, *173*(4), 353–359. https://doi.org/10.1006/jtbi.1995.0068

Guerra, N. G., Huesmann, L. R., & Spindler, A. (2003). Community Violence Exposure, Social Cognition, and Aggression Among Urban Elementary School Children. *Child Development*, *74*(5), 1561–1576. https://doi.org/10.1111/1467-8624.00623

Guo, X., Zheng, L., Wang, H., Zhu, L., Li, J., Wang, Q., Dienes, Z., & Yang, Z. (2013). Exposure to violence reduces empathetic responses to other's pain. *Brain and Cognition*, *82*(2), 187–191. https://doi.org/10.1016/j.bandc.2013.04.005

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, *24*, 92–97. https://doi.org/10.1016/j.copsyc.2018.09.001

Harris, A. C., & Madden, G. J. (2002). Delay Discounting and Performance on the Prisoner's Dilemma Game. *The Psychological Record*, *52*(4), 429–440. https://doi.org/10.1007/BF03395196

Hawkins, J. D., Herrenkohl, T. I., Farrington, D. P., Brewer, D., Catalano, R. F., Harachi, T. W., & Cothern, L. (2000). *Predictors of Youth Violence. Juvenile Justice Bulletin*. https://eric.ed.gov/?id=ED440196

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Heiphetz, L. (2019). Moral essentialism and generosity among children and adults. *Journal of Experimental Psychology: General*, *148*(12), 2077–2090. https://doi.org/10.1037/xge0000587

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Heyman, G. D., & Gelman, S. A. (2000). Beliefs about the origins of human psychological traits. *Developmental Psychology*, *36*(5), 663–678. https://doi.org/10.1037/0012-1649.36.5.663

Hill, P. F., Yi, R., Spreng, R. N., & Diana, R. A. (2017). Neural congruence between intertemporal and interpersonal self-control: Evidence from delay and social discounting. *NeuroImage*, *162*, 186–198. https://doi.org/10.1016/j.neuroimage.2017.08.071

Hirschi, T. (2004). Self-control and crime. In *Handbook of self-regulation: Research, theory, and applications* (pp. 537–552). The Guilford Press.

Hobbes, T. (1994). *Leviathan*. Hackett. (Original work published 1668)

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Holm, A. L., Berg, A., & Severinsson, E. (2009). Longing for Reconciliation: A Challenge for Women with Borderline Personality Disorder. *Issues in Mental Health Nursing*, *30*(9), 560–568. https://doi.org/10.1080/01612840902838579

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, *92*(5), 1644–1655. https://doi.org/10.1257/000282802762024700

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making. *Science*, *310*(5754), 1680–1683. https://doi.org/10.1126/science.1115327

Huesmann, L. R., & Kirwil, L. (2007). Why observing violence increases the risk of violent behavior by the observer. In *The Cambridge handbook of violent behavior and aggression* (pp. 545–570). Cambridge University Press. https://doi.org/10.1017/CBO9780511816840.029

Hughes, J. S. (2017). In a moral dilemma, choose the one you love: Impartial actors are seen as less moral than partial ones. *British Journal of Social Psychology*, n/a-n/a. https://doi.org/10.1111/bjso.12199

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting From Misfortune When Harmless Actions Are Judged to Be Morally Blameworthy. *Personality and Social Psychology Bulletin*, *38*(1), 52–62. https://doi.org/10.1177/0146167211430232

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The Effort Paradox: Effort Is Both Costly and Valued. *Trends in Cognitive Sciences*, *22*(4), 337–349. https://doi.org/10.1016/j.tics.2018.01.007

Javdani, S., Abdul-Adil, J., Suarez, L., Nichols, S. R., & Farmer, A. D. (2014). Gender Differences in the Effects of Community Violence on Mental Health Outcomes in a Sample of Low-Income Youth Receiving Psychiatric Care. *American Journal of Community Psychology*, *53*(3–4), 235–248. https://doi.org/10.1007/s10464-014-9638-2

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38. https://doi.org/10.1016/j.cognition.2015.05.006

Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution*, *28*(8), 474–481. https://doi.org/10.1016/j.tree.2013.05.014

Jones, B., & Rachlin, H. (2006). Social discounting. *Psychological Science*, *17*(4), 283–286. https://doi.org/10.1111/j.1467-9280.2006.01699.x

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663. https://doi.org/10.1073/pnas.1601280113

Jordan, J. J., & Rand, D. G. (2019). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/pspi0000186

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209. https://doi.org/10.1016/j.cognition.2014.10.005

Kameda, T., Inukai, K., Higuchi, S., Ogawa, A., Kim, H., Matsuda, T., & Sakagami, M. (2016). Rawlsian maximin rule operates as a common cognitive anchor in distributive justice and risky decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(42), 11817–11822. JSTOR.

Kant, I. (2002). *Groundwork for the Metaphysics of Morals*. Yale University Press. (Original work published 1797)

Kappes, A., Nussberger, A.-M., Siegel, J. Z., Rutledge, R. B., & Crockett, M. J. (2019). Social uncertainty is heterogeneous and sometimes valuable. *Nature Human Behaviour*, *3*(8), 764. https://doi.org/10.1038/s41562-019-0662-y

Kee, H. W. (1969). *Development, and the effects upon bargaining, of trust and suspicion* [University of British Columbia]. https://doi.org/10.14288/1.0104029

Koch, A., Imhoff, R., Unkelbach, C., Nicolas, G., Fiske, S., Terache, J., Carrier, A., & Yzerbyt, V. (2020). Groups' warmth is a personal matter: Understanding consensus on stereotype dimensions reconciles adversarial models of social evaluation. *Journal of Experimental Social Psychology*, *89*, 103995. https://doi.org/10.1016/j.jesp.2020.103995

Koch, A., Yzerbyt, V., Abele, A., Ellemers, N., & Fiske, S. T. (in press). Social evaluation: Comparing models across interpersonal, intragroup, intergroup, several-group, and many-group contexts. *Advances in Experimental Social Psychology*, *63*.

Kocher, M., Martinsson, P., Myrseth, K. O. R., & Wollbrant, C. (2017). Strong, bold, and kind: Self-control and cooperation in social dilemmas. *Experimental Economics*, *20*(1), 44–69.

Koenig, M. A., Tiberius, V., & Hamlin, J. K. (2019). Children's Judgments of Epistemic and Moral Agents: From Situations to Intentions. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *14*(3), 344–360. https://doi.org/10.1177/1745691618805452

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*(7138), 908–911. https://doi.org/10.1038/nature05631

Kouchaki, M., & Gino, F. (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences*, *113*(22), 6166–6171. https://doi.org/10.1073/pnas.1523586113

Krause, N., & Ellison, C. G. (2003). Forgiveness by God, Forgiveness of Others, and Psychological Well–Being in Late Life. *Journal for the Scientific Study of Religion*, *42*(1), 77–93. https://doi.org/10.1111/1468-5906.00162

Lamba, A., Frank, M. J., & FeldmanHall, O. (2020). Anxiety Impedes Adaptive Social Learning Under Uncertainty. *Psychological Science*, *31*(5), 592–603. https://doi.org/10.1177/0956797620910993

Lane, J. D., Wellman, H. M., & Gelman, S. A. (2013). Informants' Traits Weigh Heavily in Young Children's Trust in Testimony and in Their Epistemic Inferences. *Child Development*, *84*(4), 1253–1268. https://doi.org/10.1111/cdev.12029

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*(2), 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Levi, M., & Stoker, L. (2000). Political Trust and Trustworthiness. *Annual Review of Political Science*. https://doi.org/10.1146/annurev.polisci.3.1.475

Levine, S., Mikhail, J., & Leslie, A. M. (2018). Presumed innocent? How tacit assumptions of intentional structure shape moral judgment. *Journal of Experimental Psychology. General*, *147*(11), 1728–1747. https://doi.org/10.1037/xge0000459

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2009). Neural Representation of Subjective Value Under Risk and Ambiguity. *Journal of Neurophysiology*, *103*(2), 1036–1047. https://doi.org/10.1152/jn.00853.2009

Lockwood, P. L., Apps, M. A. J., & Chang, S. W. C. (2020). Is There a 'Social' Brain? Implementations and Algorithms. *Trends in Cognitive Sciences*, *24*(10), 802–813. https://doi.org/10.1016/j.tics.2020.06.011

Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., & Apps, M. A. J. (2017). Prosocial apathy for helping others when effort is required. *Nature Human Behaviour*, *1*(7). https://doi.org/10.1038/s41562-017-0131

Lockwood, P. L., & Klein-Flugge, M. (2020). Computational modelling of social cognition and behaviour—A reinforcement learning primer. *Social Cognitive and Affective Neuroscience*. https://academic.oup.com/scan/advance-article/doi/10.1093/scan/nsaa040/5813717

Lojowska, M., Mulckhuyse, M., Hermans, E. J., & Roelofs, K. (2019). Unconscious processing of coarse visual information during anticipatory threat. *Consciousness and Cognition*, *70*, 50–56. https://doi.org/10.1016/j.concog.2019.01.018

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lutz, D. J., & Keil, F. C. (2002). Early Understanding of the Division of Cognitive Labor. *Child Development*, *73*(4), 1073–1084. https://doi.org/10.1111/1467-8624.00458

Maltby, J., Macaskill, A., & Day, L. (2001). Failure to forgive self and others: A replication and extension of the relationship between forgiveness, personality, social desirability and general health. *Personality and Individual Differences*, *30*(5), 881–885. https://doi.org/10.1016/S0191-8869(00)00080-5

Marr, D. (1982). *Vision: A computational investigation*. Freeman.

Martijn, C., Spears, R., Van Der Pligt, J., & Jakobs, E. (1992). Negativity and positivity effects in person perception and inference: Ability versus morality. *European Journal of Social Psychology*, *22*(5), 453–463. https://doi.org/10.1002/ejsp.2420220504

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39. https://doi.org/10.3389/fnhum.2011.00039

McCullough, M. E. (2000). Forgiveness as Human Strength: Theory, Measurement, and Links to Well-Being. *Journal of Social and Clinical Psychology*, *19*(1), 43–55. https://doi.org/10.1521/jscp.2000.19.1.43

McCullough, M. E. (2008). *Beyond revenge: The evolution of the forgiveness instinct*. John Wiley & Sons.

McCullough, M. E., Pargament, K. I., & Thoresen, C. E. (2000). *Forgiveness: Theory, Research, and Practice*. Guilford Press.

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., Tolias, A. S., Cardin, J. A., & McCormick, D. A. (2015). Waking State: Rapid Variations Modulate Neural and Behavioral Responses. *Neuron*, *87*(6), 1143–1161. https://doi.org/10.1016/j.neuron.2015.09.012

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What We Owe to Family: The Impact of Special Obligations on Moral Judgment. *Psychological Science*, *31*(3), 227–242. https://doi.org/10.1177/0956797619900321

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic Value Underlies Asymmetric Updating of Impressions in the Morality and Ability Domains. *The Journal of Neuroscience*, *33*(50), 19406–19415. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2012). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, nss040. https://doi.org/10.1093/scan/nss040

Mill, J. S. (1863). *Utilitarianism*. Parker, Son, and Bourne.

Mirels, H. L., & Garrett, J. B. (1971). The Protestant Ethic as a personality variable. *Journal of Consulting and Clinical Psychology*, *36*(1), 40–44. https://doi.org/10.1037/h0030477

Moffitt, T. E., & Tank, T. K.-G. 2012 T. (2013). Childhood exposure to violence and lifelong health: Clinical intervention science and stress-biology research join forces. *Development and Psychopathology*, *25*(4pt2), 1619–1634. https://doi.org/10.1017/S0954579413000801

Molander, P. (1985). The Optimal Level of Generosity in a Selfish, Uncertain Environment. *The Journal of Conflict Resolution*, *29*(4), 611–618.

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and Cognition*, *25*, 67–76. https://doi.org/10.1016/j.concog.2014.01.009

Mullen, E., & Monin, B. (2016). Consistency Versus Licensing Effects of Past Moral Behavior. *Annual Review of Psychology*, *67*(1), 363–385. https://doi.org/10.1146/annurev-psych-010213-115120

Müller, S., & Rau, H. A. (2016). The relation of risk attitudes and other-regarding preferences: A within-subjects analysis. *European Economic Review*, *85*, 1–7. https://doi.org/10.1016/j.euroecorev.2016.02.004

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value Judgments and the True Self. *Personality and Social Psychology Bulletin*, *40*(2), 203–216. https://doi.org/10.1177/0146167213508791

Ng-Mak, D. S., Salzinger, S., Feldman, R. S., & Stueve, C. A. (2004). Pathologic Adaptation to Community Violence Among Inner-City Youth. *American Journal of Orthopsychiatry*, *74*(2), 196–208. https://doi.org/10.1037/0002-9432.74.2.196

Ng-Mak, D. S., Stueve, A., Salzinger, S., & Feldman, R. (2002). Normalization of Violence Among Inner-City Youth: A Formulation for Research. *American Journal of Orthopsychiatry*, *72*(1), 92–101. https://doi.org/10.1037/0002-9432.72.1.92

Nicol, K., Pope, M., Sprengelmeyer, R., Young, A. W., & Hall, J. (2013). Social Judgement in Borderline Personality Disorder. *PLOS ONE*, *8*(11), e73440. https://doi.org/10.1371/journal.pone.0073440

Noordewier, M. K., Scheepers, D. T., & Hilbert, L. P. (2019). Freezing in response to social threat: A replication. *Psychological Research*. https://doi.org/10.1007/s00426-019-01203-4

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, *355*(6357), 250–253. https://doi.org/10.1038/355250a0

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298. https://doi.org/10.1038/nature04131

Öhman, A. (1986). Face the Beast and Fear the Face: Animal and Social Fears as Prototypes for Evolutionary Analyses of Emotion. *Psychophysiology*, *23*(2), 123–145. https://doi.org/10.1111/j.1469-8986.1986.tb00608.x

Olivola, C. Y., & Shafir, E. (2013). The Martyrdom Effect: When Pain and Effort Increase Prosocial Contributions. *Journal of Behavioral Decision Making*, *26*(1), 91–105. https://doi.org/10.1002/bdm.767

Pagliaro, S., Brambilla, M., Sacchi, S., D'Angelo, M., & Ellemers, N. (2013). Initial Impressions Determine Behaviours: Morality Predicts the Willingness to Help Newcomers. *Journal of Business Ethics*, *117*(1), 37–44. https://doi.org/10.1007/s10551-012-1508-y

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Piazza, J., Goodwin, G. P., Rozin, P., & Royzman, E. B. (2014). When a Virtue is Not a Virtue: Conditional Virtues in Moral Evaluation. *Social Cognition*, *32*(6), 528–558. https://doi.org/10.1521/soco.2014.32.6.528

Pratt, T. C., & Cullen, F. T. (2000). The Empirical Status of Gottfredson and Hirschi's General Theory of Crime: A Meta-Analysis. *Criminology*, *38*(3), 931–964. https://doi.org/10.1111/j.1745-9125.2000.tb00911.x

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of approach- and avoidance-related social information. *Journal of Personality and Social Psychology*, *61*(3), 380–391.

Preißler, S., Dziobek, I., Ritter, K., Heekeren, H. R., & Roepke, S. (2010). Social Cognition in Borderline Personality Disorder: Evidence for Disturbed Recognition of the Emotions, Thoughts, and Intentions of others. *Frontiers in Behavioral Neuroscience*, *4*. https://doi.org/10.3389/fnbeh.2010.00182

Pronin, E., Olivola, C. Y., & Kennedy, K. A. (2008). Doing unto future selves as you would do unto others: Psychological distance and decision making. *Personality and Social Psychology Bulletin*, *34*(2), 224–236. https://doi.org/10.1177/0146167207310023

Rachlin, H. (2002). Altruism and selfishness. *Behavioral and Brain Sciences*, *25*(2), 239–296. https://doi.org/10.1017/S0140525X02000055

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75. https://doi.org/10.1037/a0021867

Rand, D. G., Ohtsuki, H., & Nowak, M. A. (2009). Direct reciprocity with costly punishment: Generous tit-for-tat prevails. *Journal of Theoretical Biology*, *256*(1), 45–57. https://doi.org/10.1016/j.jtbi.2008.09.015

Rawls, J. (1971). *A Theory of Justice*. Belknap Press of Harvard University Press.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61–79. https://doi.org/10.1037/0033-295X.86.1.61

Reeder, G. D., & Coovert, M. D. (1986). Revising an Impression of Morality. *Social Cognition*, *4*(1), 1–17. https://doi.org/10.1521/soco.1986.4.1.1

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*, 13289. https://doi.org/10.1038/ncomms13289

Richard, Dawkins, John, & Krebs, R. (2005). *Chapter 10 Animal Signals: Information or Manipulation ?* /paper/Chapter-10-Animal-Signals-%3A-Information-or-Richard-Dawkins/67392ce3e0a993e99e1a61111a08c5f553d568ad

Righetti, F., & Finkenauer, C. (2011). If You Are Able to Control Yourself, I Will Trust You: The Role of Perceived Self-Control in Interpersonal Trust. *Journal of Personality and Social Psychology*, *100*(5), 874–886. https://doi.org/10.1037/a0021827

Riskey, D. R., & Birnbaum, M. H. (1974). Compensatory effects in moral judgment: Two rights don't make up for a wrong. *Journal of Experimental Psychology*, *103*(1), 171–173. https://doi.org/10.1037/h0036892

Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society of London B: Biological Sciences*, *265*(1394), 427–431. https://doi.org/10.1098/rspb.1998.0312

Robinson, O. J., Vytal, K., Cornwell, B. R., & Grillon, C. (2013). The impact of anxiety upon cognition: Perspectives from human threat of shock studies. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00203

Roelofs, K., Hagenaars, M. A., & Stins, J. (2010). Facing Freeze: Social Threat Induces Bodily Freeze in Humans. *Psychological Science*, *21*(11), 1575–1581. https://doi.org/10.1177/0956797610384746

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, *74*, 24–37. https://doi.org/10.1016/j.jesp.2017.08.003

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology*, *69*, 44–58. https://doi.org/10.1016/j.jesp.2016.09.007

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Sacco, D. F., Brown, M., Lustgraaf, C. J. N., & Hugenberg, K. (2017). The Adaptive Utility of Deontology: Deontological Moral Decision-Making Fosters Perceptions of Trust and Likeability. *Evolutionary Psychological Science*, *3*(2), 125–132. https://doi.org/10.1007/s40806-016-0080-6

Sandage, S. J., Long, B., Moen, R., Jankowski, P. J., Worthington, E. L., Wade, N. G., & Rye, M. S. (2015). Forgiveness in the Treatment of Borderline Personality Disorder: A Quasi-Experimental Study. *Journal of Clinical Psychology*, *71*(7), 625–640. https://doi.org/10.1002/jclp.22185

Sansone, R. A., Kelley, A. R., & Forbis, J. S. (2013). The Relationship Between Forgiveness and Borderline Personality Symptomatology. *Journal of Religion and Health*, *52*(3), 974–980. https://doi.org/10.1007/s10943-013-9704-3

Saucet, C., & Villeval, M. C. (2019). Motivated memory in dictator games. *Games and Economic Behavior*, *117*, 250–275. https://doi.org/10.1016/j.geb.2019.05.011

Scanlon, T. M. (1998). *What we owe to each other* (Vol. 66). Belknap Press of Harvard University Press.

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology*, *84*(1), 60–79. https://doi.org/10.1037/0022-3514.84.1.60

Shaw, A. (2013). Beyond "to Share or Not to Share": The Impartiality Account of Fairness. *Current Directions in Psychological Science*, *22*(5), 413–417. https://doi.org/10.1177/0963721413484467

Siegel, J. Z., Curwell-Parry, O., Pearce, S., Saunders, K. E. A., & Crockett, M. J. (2020). A Computational Phenotype of Disrupted Moral Inference in Borderline Personality Disorder. *Biological*

Psychiatry: Cognitive Neuroscience and Neuroimaging. https://doi.org/10.1016/j.bpsc.2020.07.013

Siegel, J. Z., Estrada, S., Crockett, M. J., & Baskin-Sommers, A. (2019). Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males. *Nature Communications*, *10*(1), 1942. https://doi.org/10.1038/s41467-019-09962-9

Siegel, J. Z., Mathys, C., Rutledge, R., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*.

Simone, C. D. (n.d.). *What is a good question asker better at? From no generalization, to overgeneralization, to adults-like selectivity across childhood*. 7.

Singer, P. (1993). *Practical Ethics*. Cambridge University Press.

Skowronski, J. J., & Carlston, D. E. (1987). Social Judgment and Social Memory: The Role of Cue Diagnosticity in Negativity, Positivity, and Extremity Biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Skowronski, J. J., & Carlston, D. E. (1992). Caught in the act: When impressions based on highly diagnostic behaviours are resistant to contradiction. *European Journal of Social Psychology*, *22*(5), 435–452. https://doi.org/10.1002/ejsp.2420220503

Snyder, C. R., & Lopez, S. J. (2001). *Handbook of Positive Psychology*. Oxford University Press.

Soutschek, A., Ruff, C. C., Strombach, T., Kalenscher, T., & Tobler, P. N. (2016). Brain stimulation reveals crucial role of overcoming self-centeredness in self-control. *Science Advances*, *2*(10). https://doi.org/10.1126/sciadv.1600992

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. https://doi.org/10.1073/pnas.1014345108

Starmans, C., & Bloom, P. (2016). When the spirit is willing, but the flesh is weak: Developmental differences in judgments about inner moral conflict. *Psychological Science*, *27*(11), 1498–1506. https://doi.org/10.1177/0956797616665813

Stevens, J. R., Cushman, F. A., & Hauser, M. D. (2005). Evolving the Psychological Mechanisms for Cooperation. *Annual Review of Ecology, Evolution, and Systematics*, *36*, 499–518.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Thielmann, I., Hilbig, B. E., & Niedtfeld, I. (2014). Willing to Give but Not to Forgive: Borderline Personality Features and Cooperative Behavior. *Journal of Personality Disorders*, *28*(6), 778–795. https://doi.org/10.1521/pedi_2014_28_135

Todorov, A., & Oh, D. (in press). The Structure and Perceptual Basis of Social Judgments from Faces. *Advances in Experimental Social Psychology*, *63*.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating Faces on Trustworthiness After Minimal Time Exposure. *Social Cognition*, *27*(6), 813–833. https://doi.org/10.1521/soco.2009.27.6.813

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology*, *39*(6), 549–562. https://doi.org/10.1016/S0022-1031(03)00059-3

Tsang, J.-A. (2002). Moral Rationalization and the Integration of Situational Factors and Psychological Processes in Immoral Behavior. *Review of General Psychology*, *6*(1), 25–50. https://doi.org/10.1037/1089-2680.6.1.25

Tyler, T. R. (2000). Social Justice: Outcome and Procedure. *International Journal of Psychology*, *35*(2), 117–125. https://doi.org/10.1080/002075900399411

Udow-Phillips, M., & Lantz, P. (2020). Trust in Public Health Is Essential Amid the COVID-19 Pandemic. *Journal of Hospital Medicine*, *15*(7). https://doi.org/10.12788/jhm.3474

Unkelbach, C., Alves, H., & Koch, A. (2020). Negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. *Advances in Experimental Social Psychology*, *62*.

Unoka, Z., Fogd, D., Füzy, M., & Csukly, G. (2011). Misreading the facial signs: Specific impairments and error patterns in recognition of facial emotions with negative valence in borderline personality disorder. *Psychiatry Research*, *189*(3), 419–425. https://doi.org/10.1016/j.psychres.2011.02.010

Unoka, Z., Seres, I., Áspán, N., Bódi, N., & Kéri, S. (2009). Trust Game Reveals Restricted Interpersonal Transactions in Patients With Borderline Personality Disorder. *Journal of Personality Disorders*, *23*(4), 399–409. https://doi.org/10.1521/pedi.2009.23.4.399

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. https://doi.org/10.1037/0033-2909.134.3.383

van ’t Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*(3), 796–803. https://doi.org/10.1016/j.cognition.2008.07.002

Vives, M.-L., & FeldmanHall, O. (2018). Tolerance to ambiguous uncertainty predicts prosocial behavior. *Nature Communications*, *9*(1), 2156. https://doi.org/10.1038/s41467-018-04631-9

Vonasch, A. J., Reynolds, T., Winegard, B. M., & Baumeister, R. F. (2018). Death Before Dishonor: Incurring Costs to Protect Moral Reputation. *Social Psychological and Personality Science*, *9*(5), 604–613. https://doi.org/10.1177/1948550617720271

Vonasch, A., & Sjåstad, H. (2019). *Future-orientation (as trait and state) promotes reputation-protective choice in moral dilemmas*. PsyArXiv. https://doi.org/10.31234/osf.io/x2afb

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266–281. https://doi.org/10.1016/j.cognition.2016.11.007

W.D. Ross. (1930). *The Right and the Good*. Oxford University Press.

Weber, M. (2002). *The Protestant Ethic and the “spirit” of Capitalism and Other Writings*. Penguin.

Whiteley, S. (2004). The Evolution of the Therapeutic Community. *Psychiatric Quarterly*, *75*(3), 233–248. https://doi.org/10.1023/B:PSAQ.0000031794.82674.e8

Willis, J., & Todorov, A. (2006). First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Witvliet, C. van O., Ludwig, T. E., & Laan, K. L. V. (2001). Granting Forgiveness or Harboring Grudges: Implications for Emotion, Physiology, and Health. *Psychological Science*, *12*(2), 117–123. https://doi.org/10.1111/1467-9280.00320

Wojciszke, B. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology*, *16*(1), 155–188. https://doi.org/10.1080/10463280500229619

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the Dominance of Moral Categories in Impression Formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001

Worthington, E. L., & Scherer, M. (2004). Forgiveness is an emotion-focused coping strategy that can reduce health risks and promote health resilience: Theory, review, and hypotheses. *Psychology & Health*, *19*(3), 385–405. https://doi.org/10.1080/0887044042000196674

Wu, J., & Axelrod, R. (1995). How to Cope with Noise in the Iterated Prisoner's Dilemma. *The Journal of Conflict Resolution*, *39*(1), 183–189.

Xiong, W., Gao, X., He, Z., Yu, H., Liu, H., & Zhou, X. (2020). Affective evaluation of others' altruistic decisions under risk and ambiguity. *NeuroImage*, *218*, 116996. https://doi.org/10.1016/j.neuroimage.2020.116996

Ybarra, O., Chan, E., & Park, D. (2001). Young and Old Adults' Concerns About Morality and Competence. *Motivation and Emotion*, *25*(2), 85–100. https://doi.org/10.1023/A:1010633908298

Yi, R., Johnson, M. W., & Bickel, W. K. (2005). Relationship between cooperation in an iterated prisoner's dilemma game and the discounting of hypothetical outcomes. *Learning & Behavior*, *33*(3), 324–336. https://doi.org/10.3758/BF03192861

Yu, A., & Dayan, P. (2003). Expected and unexpected uncertainty: ACh and NE in the neocortex. *Advances in Neural Information Processing Systems*. http://discovery.ucl.ac.uk/185399/

Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2019). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/bul0000218

Zhong, C.-B., Liljenquist, K., & Cain, D. M. (2009). Moral self-regulation: Licensing and compensation. In *Psychological perspectives on ethical behavior and decision making* (pp. 75–89). Information Age Publishing, Inc.