TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 11 (2019) 409–432

ISSN:1756-8765 online
DOI: 10.1111/tops.12382

This article is part of the topic "Computational Approaches to Social Cognition," Samuel Gershman and Fiery Cushman (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

Modeling Morality in 3-D: Decision-Making, Judgment, and Inference

Hongbo Yu,[a,b] Jenifer Z. Siegel,[a,b] Molly J. Crockett[a,b]

[a]*Department of Experimental Psychology, University of Oxford*
[b]*Department of Psychology, Yale University*

Received 15 September 2017; received in revised form 9 July 2018; accepted 13 July 2018

Abstract

Humans face a fundamental challenge of how to balance selfish interests against moral considerations. Such trade-offs are implicit in moral *decisions* about what to do; *judgments* of whether an action is morally right or wrong; and *inferences* about the moral character of others. To date, these three dimensions of moral cognition–decision-making, judgment, and inference–have been studied largely independently, using very different experimental paradigms. However, important aspects of moral cognition occur at the intersection of multiple dimensions; for instance, moral hypocrisy can be conceived as a disconnect between moral decisions and moral judgments. Here we describe the advantages of investigating these three dimensions of moral cognition within a single computational framework. A core component of this framework is *harm aversion*, a moral sentiment defined as a distaste for harming others. The framework integrates economic utility models of harm aversion with Bayesian reinforcement learning models describing beliefs about others' harm aversion. We show how this framework can provide novel insights into the mechanisms of moral decision-making, judgment, and inference.

Keywords: Moral cognition; Moral decision-making; Moral judgment; Moral inference; Computational models; Harm aversion

Correspondence should be sent to Molly J. Crockett, Department of Psychology, Yale University, New Haven, CT 06520. E-mail: molly.crockett@yale.edu

1. Introduction

On December 14, 2012, Sandy Hook Elementary School teacher Victoria Soto "threw herself in front of her first-grade students" to protect them from a gunman attacking the school (Planas, 2012). This tragic story illustrates three key dimensions of moral cognition: Soto made a *moral decision* to put her own life in danger to protect her students; you, the reader, probably made a *moral judgment* about whether Soto did the right thing; and from there you probably made a further *moral inference* about what kind of person Soto is in general. Research in moral psychology has traditionally investigated moral cognition along these same three dimensions (Fig. 1):

1. Moral decision-making: how people make decisions that affect the welfare of others (e.g., Batson, Duncan, Ackerman, Buckley, & Birch, 1981; Batson, Fultz, & Schoenrade, 1987; FeldmanHall, Dalgleish, Evans, & Mobbs, 2015; Gao et al., 2018; Garrett, Lazzaro, Ariely, & Sharot, 2016; Greene & Paxton, 2009; Hsu, Anen, & Quartz, 2008; Koenigs et al., 2007; Rand, Greene, & Nowak, 2012; Sáez et al., 2015: Shalvi, Gino, Barkan, & Ayal, 2015; Zhu et al., 2014).
2. Moral judgment: how people make judgments about the moral appropriateness of actions and assign blame and punishment, or praise and reward (e.g., Baron, 2014; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Malle, Guglielmo, & Monroe, 2014; Schein & Gray, 2014, 2015, 2018; Shenhav & Greene, 2010; Wojciszke, Parzuchowski, & Bocian, 2015; Young & Saxe, 2008; Young et al., 2010).
3. Moral inference: how people form beliefs about the moral character of agents based on observations of morally relevant behaviors (e.g., Alicke & Zell, 2009; Bostyn &

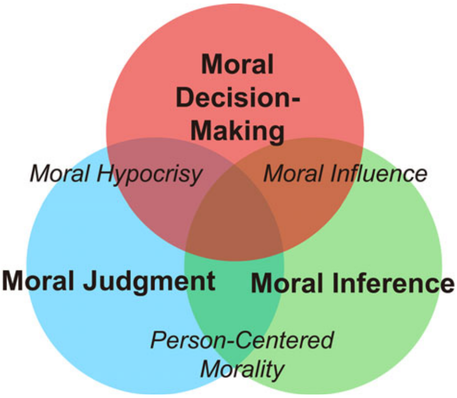

Fig. 1. Three dimensions of moral cognition. Most moral cognition research investigates one of three dimensions: moral decision-making, moral judgment, or moral inference. Important phenomena lie at the intersection of two or more dimensions. For example, moral hypocrisy can be conceptualized as a disconnect between moral decision-making and moral judgment, where hypocrites judge others harshly for the same decisions they make themselves; in moral influence, inferences about the moral character of others shape one's own moral decisions; and work on person-centered morality demonstrates that inferences about moral character spill over into moral judgments of individual actions.

Roets, 2017; Behrens, Hunt, Woolrich, & Rushworth, 2008; Diaconescu et al., 2014; Everett, Pizarro, & Crockett, 2016; Hackel, Doll, & Amodio, 2015; Giner-Sorolla & Chapman, 2017; Kliemann, Young, Scholz, & Saxe, 2008; Kleiman-Weiner, Saxe, & Tenenbaum, 2017; Knobe, 2010; Nadler, 2012).

To date, these three dimensions have been investigated mostly independently, usually by different researchers using very different experimental paradigms. For example, moral decision-making has typically been studied with tasks involving incentivized choices affecting the welfare of others; moral judgment has typically been studied using hypothetical "dilemma" scenarios; and moral inference has typically been studied using narrative descriptions of moral/immoral behaviors. Here, we advocate for investigating moral decision-making, judgment, and inference within the same experimental framework that incorporates computational models of cognitive processes. We propose that this approach can advance the study of moral cognition in several ways. First, it can reveal common computations underlying moral decision-making, judgment, and inference. Second, it can facilitate the investigation of many important moral phenomena that involve intersections across dimensions, such as moral hypocrisy, moral influence, and person-centered moral judgments (Fig. 1).

In the following, we introduce an example experimental framework that can be used to concurrently investigate three dimensions of moral cognition in the domain of harm. This framework incorporates computational models that describe how external features of a moral problem (e.g., harm, benefit, causation, intention, character, etc.) can be transformed into an internal utility, and how this utility is used to guide moral decision-making, judgment, and inference (Crockett, 2016). Formal model comparison procedures are used to compare the predictive power of different models that make different assumptions about how people make decisions, judgments, and inferences, testing the ability of a hypothesized set of cognitive processes to account for the entire set of choices people make as well as patterns of brain activity (Daw, 2011; Fehr & Krajbich, 2014; Hutcherson et al., 2015; Konovalov et al., 2018; Krakauer et al., 2017; Love, 2015 O'Doherty, Hampton, & Kim, 2007). Using this approach, it may be possible to reveal common computations in moral decision-making, judgment, and inference by examining, for example whether similar models can describe behavior along different dimensions; whether individual differences in one dimension predict individual differences in other dimensions; and whether there are similar neural processes underlying different computations across dimensions. We provide examples of such evidence in the following sections.

Thereafter, we describe how this approach may be able to illuminate the nature of complex moral phenomena that lie at the boundary of two areas of moral cognition: person-centered morality (Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Knobe, 2010; Tannenbaum, Uhlmann, & Diermeier, 2011; Uhlmann, Pizarro, & Diermeier, 2015), moral hypocrisy (Batson, Kobrynowicz, Dinnerstein, Kampf, & Wilson, 1997; Gino, Norton, & Weber, 2016; Graham, Meindl, Koleva, Iyer, & Johnson, 2015; Sharma, Mazar, Alter, & Ariely, 2014; Szabados & Soifer, 2004), and moral influence (Bandura, 1969; Cialdini & Goldstein, 2004; Hoffman, 1970; Gino, Ayal, & Ariely, 2009; Macaulay

& Berkowitz, 1970; Staub, 1971). Although the illustrative examples provided in this paper are specific to just one domain of morality (i.e., harm), the approach we describe can potentially be applied to other moral domains as well.

2. Harm aversion as a core component of moral cognition across dimensions

Our framework adopts the view that the computation of utility or value of a particular action for oneself, other individuals, and/or society comprises a core component of moral cognition (Bartels, Bauman, Cushman, Pizarro, & McGraw, 2015; Crockett, 2013, 2016; Cushman, 2013; Shenhav & Greene, 2010) and is related to the moral philosophy of utilitarianism, which posits that morally right action is the action that produces the most good or utility (e.g., Mill, 1863/1998). We propose that a key subcomponent of utility in moral cognition is *harm aversion*: a moral sentiment defined as a distaste for harming others. Although it is still debated whether harm is the essence of morality (Gray, Young, & Waytz, 2012; Schein & Gray, 2018) or just one of several moral "foundations" (Graham et al., 2011), it is widely acknowledged that avoiding harm to others is a universal moral principle (Gert, 2004; Keane, 2015) and comprises the majority of moral experiences in daily life (Hofmann, Wisneski, Brandt, & Skitka, 2014).

Studies of harm aversion in moral judgment, decision-making, and inference have typically relied on very different methods. Most studies of moral judgment rely on hypothetical scenarios, such as the classic "trolley problem" where participants are asked if it's acceptable to push a large man off a bridge to stop a trolley from running over several track workers (e.g., Greene et al., 2001). Meanwhile, studies of moral decision-making generally ask participants to make choices in the laboratory that have actual consequences for themselves and others, such as trading off money for oneself against electric shocks to others (e.g., FeldmanHall et al., 2012; Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014). Finally, research on moral inference typically asks participants to form impressions of others based on descriptions of morally relevant behaviors, such as performing good deeds or committing crimes (e.g., Goodwin, Piazza, & Rozin, 2014; Uhlmann et al., 2015).

The methods used to study each of these dimensions of harm-based moral cognition have been developed for good reasons, but using diverse paradigms to measure different dimensions may hinder the identification of common computational processes that operate across multiple dimensions. If such common computations exist, individual variability in one dimension of moral cognition might predict variability along other dimensions. For example, individuals who are highly harm averse in their moral decisions may also be highly harm averse in their moral judgments. It is difficult to address this question definitively using different paradigms to measure different dimensions. For example, it is difficult to know whether the same kind of harm aversion motivates the judgment that one should not push people off bridges as well as the decision to avoid delivering electric shocks to others. If one observes a positive correlation between harm aversion in trolley judgments and shock decisions, it is difficult to attribute this relationship to a common computational process because trolley problems do not explicitly measure computation; if no correlation is observed, this may be due to the large differences between paradigms. One way to make meaningful comparisons between different dimensions of moral cognition is to develop a paradigm that can simultaneously interrogate the computational processes underlying moral decision-making, judgment, and inference within the same setting. This approach makes it possible to begin testing the hypothesis that different dimensions of moral cognition are built upon a few basic computations, such as computing utility by trading off costs and benefits to oneself against costs and benefits to others (see also Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016).

2.1. *Modeling harm aversion in moral decision-making*

Recently, we examined harm aversion in moral decision-making by investigating how people trade off money for themselves against pain for themselves and others (Crockett et al., 2014, 2015; Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017; Fig. 2). In this paradigm, participants ("Decider") make choices between different amounts of money and different numbers of painful electric shocks directed toward either themselves or an anonymous other person ("Receiver"). Computational models formally quantify the relative values people ascribe to pain for themselves and others, and how those values are transformed into choices.

The valuation of harmful actions affecting oneself and others can be described by different models that contain different parameters or have different ways of integrating the values of pain and money. Crockett et al. (2014) compared a number of models and found participants' decisions were best described by a model that contained independent parameters describing harm aversion for self and others:

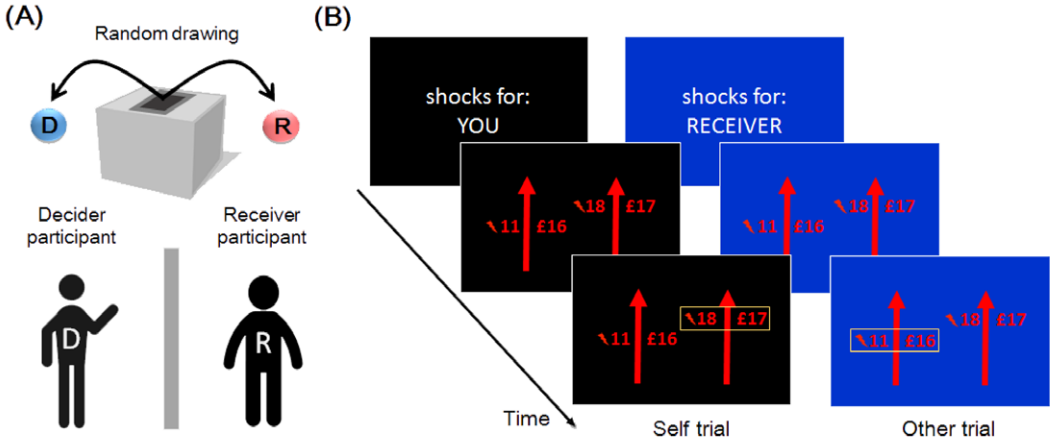

Fig. 2. Moral decision task. (A) Participants are randomly assigned to the roles of Decider and Receiver. (B) Deciders make a series of choices where they are asked to trade-off money for themselves against pain to either themselves (Self trial) or the Receiver (Other trial).

$$V(\text{harm}) = (1 - \kappa)\Delta m - \kappa\Delta s \tag{1a}$$

$$P(\text{choose harm}) = \frac{1}{1 + e^{-\beta V(\text{harm})}} \tag{1b}$$

The model relates objective features of the choice options (here, amounts of money difference, Δm, and shock difference, Δs) to their underlying subjective values (i.e., *V*(harm)). A softmax function (Eq. 1b) transforms the relative subjective value of choosing the harmful option into a probability of that choice, where the parameter β defines the steepness of the slope in the softmax function: Larger β corresponds to a steeper slope and indicates more deterministic choice preference. The money and shock terms are scaled by a harm aversion parameter (κ) that quantifies the exchange rate between money and pain and takes on different values for pain to self and others. Strikingly, across several studies harm aversion for others was consistently greater on average than harm aversion for self (Crockett et al., 2014, 2015, 2017), an effect that has been replicated by an independent research group using a different paradigm (Volz, Welborn, Gobel, Gazzaniga, & Grafton, 2017). People were willing to pay more to prevent shocks to others than to themselves, and required more compensation to increase shocks to others than themselves; that is, their behavior was “hyperaltruistic” (Kitcher, 1993). This pattern of choice cannot be readily explained by classic theories of empathy or social preferences, which posit that people value others’ welfare no more (and often much less) than their own welfare (Batson et al., 1981; Engel, 2011; Singer et al., 2004). However, hyperaltruism is consistent with work in moral psychology suggesting people experience aversive feelings (e.g., guilt or fear of blame) when causing bad outcomes, especially when those outcomes affect others (Cushman, Gray, Gaffey, & Mendes, 2012a; Ritov & Baron, 1990). These aversive feelings, or expectations of them, might degrade the value of actions that harm others (Baumeister, Stillwell, & Heatherton, 1994; Chang, Smith, Dufwenberg, & Sanfey, 2011; Charness & Dufwenberg, 2006; Lewis, 1971; Yu, Hu, Hu, & Zhou, 2014).

There are at least two possible mechanistic explanations for hyperaltruism. First, people may compute the value of others’ pain as more aversive than their own pain. Alternatively, money gained immorally (i.e., via harming others) may be subjectively less valuable than money gained from harming only oneself, perhaps due to discomfort, guilt, or anticipation of being blamed or judged. Because the harm aversion parameter in the model represents an exchange rate between money and pain, the parameter estimates alone do not straightforwardly reveal the underlying cognitive process. However, because harm aversion is the output of a computational process integrating the values of money and pain, it is reasonable to hypothesize that harm aversion might covary with neural responses to money, pain, or both. Thus, by combining the model with fMRI it is possible to interrogate how the brain represents profit and pain during moral decision-making, and whether individual differences in neural responses to profit or pain track with individual differences in hyperaltruism. This approach sidesteps the necessity for informal reverse inference because the model makes trial-by-trial predictions about neural responses to profit, pain, value, and so on (Behrens, Hunt, & Rushworth, 2009). If hyperaltruism arises from an increased weighting of others' pain relative to one's own, then pain-sensitive brain regions should show relatively increased responses to others' pain, to the extent people are hyperaltruistic. Meanwhile if hyperaltruism arises from a reduced valuation of ill-gotten gains relative to profits gained morally, then profit-sensitive brain regions should show relatively reduced responses to ill-gotten gains, to the extent people are hyperaltruistic.

A recent neuroimaging study (Crockett et al., 2017) found strong evidence for the latter hypothesis. Although the insula and anterior cingulate cortex (ACC) responded to anticipated pain for self (and to a lesser extent for others), individual differences in "empathic" pain responses in these regions did not predict individual differences in hyperaltruism. In fact, there were no brain areas where differential responses to others' versus own pain correlated with hyperaltruism. Meanwhile, responses in the brain's valuation network, in particular the dorsal striatum (DS), showed reduced responses to money gained from shocking others relative to money gained from shocking self, to the extent that people were hyperaltruistic (Fig. 3). This indicates that moral behavior might arise from a devaluation of profits gained from harming others.

2.2. Modeling harm aversion in moral judgment

The moral decision task can be easily modified to investigate moral judgment. In the moral judgment task, instead of deciding whether to profit from inflicting pain on others, participants are presented with decisions that others have made and asked to judge the extent to which those decisions are blameworthy or praiseworthy (Fig. 4A). Computational models can then be built to describe how judgments of blame and praise are sensitive to, for example the amount of pain inflicted, the amount of profit gained, and whether the decision was made actively or passively. Crucially, by asking participants to complete both the moral decision task and the moral judgment task, it is possible to

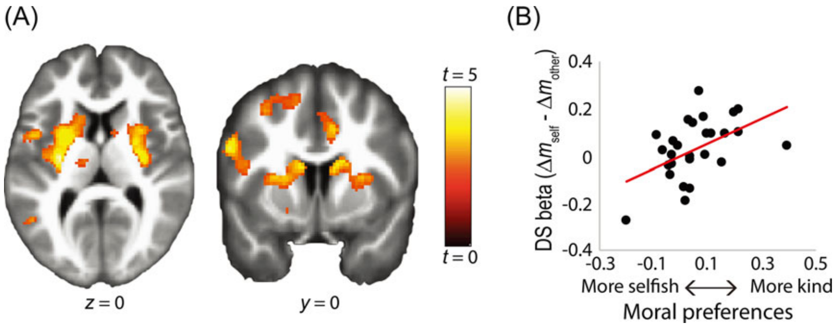

Fig. 3. The neural basis of hyperaltruism. (A and B) In bilateral DS, reduced responses to profits gained from harming others correlated with hyperaltruism ($\kappa_{\text{other}} - \kappa_{\text{self}}$). Figure adapted from Crockett et al. (2017).

describe how one's own moral preferences are related to moral judgments of others' behavior within an identical context. In doing so it is possible to address questions such as whether "conscience" in moral decision-making can be conceptualized as a turning-inward of moral judgments normally applied to others (Smith, 1759/2002; Freud, 1961), and whether people who are more harm averse in moral decision-making are also more harm averse in their moral judgments.

In a recent study (Crockett et al., 2017), participants first completed the moral decision task depicted in Fig. 2, through which we could estimate their degree of harm aversion for themselves (κ_s) and for others (κ_o). Next, participants completed the moral judgment task depicted in Fig. 4A where on each trial they judged the blameworthiness of choices made by another agent. Crucially, in the moral judgment task the differences in shocks (Δs) and profit (Δm) resulting from the agent's choices were decorrelated across trials. This not only enabled us to estimate the independent contributions of pain and profits to moral judgments, but also how people's own harm aversion influences their reliance on pain and profit in making moral judgments. Thus, in our moral judgment model, trial-by-trial blame judgments were regressed against trial features (Δs and Δm), individuals' own harm aversion (κ_s and κ_o), and their interactions:

$$\begin{aligned} Blame_t &= \beta_0 + \beta_1 \Delta m + \beta_2 \Delta s + \beta_3 \Delta m \kappa_o + \beta_4 \Delta m \kappa_s \\ &+ \beta_5 \Delta s \kappa_o + \beta_6 \Delta s \kappa_s + \beta_7 \Delta m \kappa_s \kappa_o + \beta_8 \Delta s \kappa_s \kappa_o \end{aligned} \tag{2}$$

The regression revealed several key findings. First, moral judgments of blameworthiness involved very similar computations as in moral decision-making. Blame was negatively correlated with the additional profit (Δm) caused by choosing the more harmful option and positively correlated with the additional pain (Δs) caused by choosing the more harmful option. This indicates that although blame is sensitive to harmful outcomes (Crockett, Clark, Hauser, & Robbins, 2010; Cushman, Young, & Hauser, 2006), the profits gained seem to justify harm at least partially, consistent with work using hypothetical

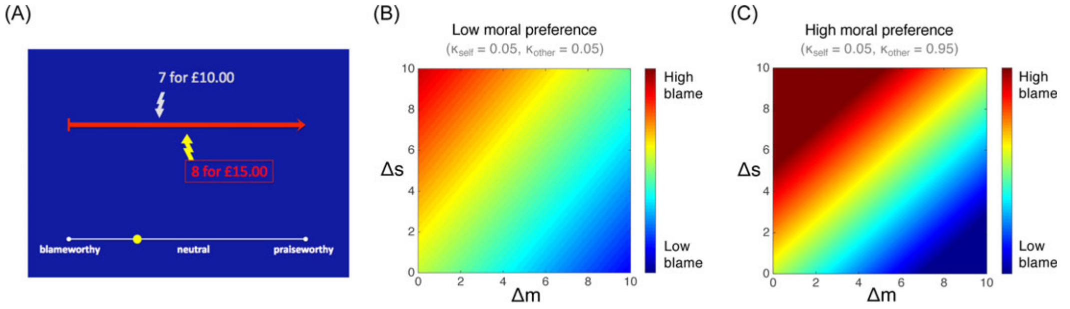

Fig. 4. Moral judgment task and model. (A) Participants are presented with decisions that others have made and asked to judge the extent to which those decisions are blameworthy or praiseworthy. (B and C) Heatmaps depicting patterns of moral judgment by participants with low hyperaltruism (B) and high hyperaltruism (C) vary as a function of extra shocks and extra money resulted from choosing the more harmful option. Figure adapted from Crockett et al. (2017).

scenarios (Xie, Yu, Zhou, Sedikides, & Vohs, 2014). Notably, these results show that money and pain exerted opposing effects on the computation of blame in moral judgments, just as they exerted opposing effects on the computation of value in moral decision-making.

Second, participants' own harm aversion preferences (κ_s and κ_o) modulated the influence of profit and pain on blame, such that the participants who themselves were more averse to causing pain to others (higher κ_o) made more extreme blame judgments and cared more about harm and less about profit in making blame judgments (Fig. 4B and C). In other words, people who were more harm averse in moral decision-making also showed a stronger influence of harm on moral judgments.

In addition, data from the moral judgment task allowed us to probe a long-standing theory about the nature of moral preferences (Smith, 1759/2002). Crockett et al. (2017) hypothesized that the devaluation of ill-gotten gains could arise from a top-down modulation of value-sensitive regions by lateral prefrontal cortical (LPFC) regions involved in representing moral norms (Buckholtz, 2015), similar to the way long-term health goals represented in LPFC modulate the value of unhealthy foods (Hare, Camerer, & Rangel, 2009). That is, LPFC may represent shared norms about what is morally appropriate and modulate the value of profits to the extent they are gained via blameworthy actions. This account predicts that LPFC activity during moral decision-making tracks the blameworthiness of harmful actions. To test this account, we used our model of moral judgment (Eq. 2) to construct an individualized "blame regressor" for each participant who completed the moral decision task in the fMRI scanner and probed the relationship between LPFC activity at time of choice and blameworthiness of each choice. Note that the participants in the fMRI scanner never made blame judgments themselves. Rather, our goal was to predict their LPFC activity based on how other participants judged the blameworthiness of harmful actions. Such analysis was possible because both moral decision-making and moral judgment were measured quantitatively within the same experimental paradigm and computational framework. If brain activity during moral decision-making could be predicted by a model of moral judgments, this would provide support for the hypothesis that common computations underlie moral judgment and moral decision-making.

Remarkably, LPFC activity at time of choice was indeed significantly correlated with other people's moral judgments of the blameworthiness of choosing the profitable but harmful option. Its response was strongest on those trials where profiting through harm was judged to be the most blameworthy by others–prototypically when the harmful action inflicted a large amount of pain for a tiny amount of profit. This finding captures the essence of how moral norms operate: When we make moral decisions, we simulate how an "impartial spectator" would judge us for violating the norm. Finally, supporting a moral devaluation account, we found that during moral decisions, LPFC was functionally connected with the same region of striatum that showed a reduced response to ill-gotten gains. These findings suggest that rather than restraining self-interest via inhibitory control processes, moral norms modulate the value of harmful actions. In other words, it's not that people are constantly tempted to harm others for their own benefit and have to override these temptations. Rather, moral norms make selfish actions less tempting in the first place.

While the above paradigm focused on moral judgments by disinterested third-parties, it can also be adapted to investigate moral judgments and affective responses of second-party targets of moral decisions. For example, in a recent fMRI study of gratitude (Yu, Gao, Zhou, & Zhou, 2018), participants received costly help from a co-player, who sacrificed their own profits to reduce participants' pain. Gratitude was well predicted by a model that integrated the co-player's sacrificed profits and participants' pain reduction in a way remarkably similar to the models of moral decision-making and judgment described above. Specifically, gratitude was positively related to pain reduction and the co-player's sacrificed profits. Moreover, trial-by-trial gratitude as predicted by the model was correlated with activity in vmPFC, just as trial-by-trial estimates of subjective value in moral decision-making correlated with responses in this region (Crockett et al., 2017). Finally, results indicated that gratitude was more sensitive to the co-player's sacrificed profits than pain reduction just as moral decisions were explained better by neural responses to profits than pain in a similar setting. Together these findings suggest that basic computations of utility integrating costs and benefits for self and others may contribute to multiple dimensions of moral cognition, including decisions to harm others, evaluations of blame and praise, and feelings of gratitude.

2.3. Modeling harm aversion in moral inference

Moral evaluation does not stop at the level of judging single events (e.g., right vs. wrong), but very often proceeds from there to making inferences about the moral character of agents (e.g., good vs. evil), a process that has been referred to as moral inference (cf. Everett et al., 2016, 2018; Helzer & Critcher, 2018; Knobe, 2010; Uhlmann et al., 2015). Accurately inferring the moral character of others helps predict their behaviors (Fiske, Cuddy, & Glick, 2007). Some forms of moral inference might be understood as a dynamic, evidence-accumulation process. However, the cognitive mechanisms through which people form and update beliefs about the moral character of others are not well understood.

We adapted our experimental paradigm to study the computational processes guiding moral inference (Fig. 5). In particular, we were interested in testing the possibility that people would successfully predict the moral decisions of other agents by accurately inferring their level of harm aversion and updating beliefs about harm aversion in accordance with Bayes' rule, which would be consistent with recent work on inference in the domains of perception (Ma, Beck, Latham, & Pouget, 2006), economic value (Schwartenbeck et al., 2015), and social intention (Behrens et al., 2009; Diaconescu et al., 2014, 2017). In the moral inference task, instead of deciding whether to profit from inflicting pain on others, participants are asked to predict the decisions that other agents will make (Fig. 5). Computational models describe how beliefs about the harm aversion of other agents develop over time. By examining moral inference and moral decision-making within the same framework, it is possible to test whether people use similar computations to make moral decisions themselves and predict the moral decisions of others.

In a series of studies (Siegel, Mathys, Rutledge, & Crockett, in press), participants predicted sequences of choices in the moral decision task made by two agents. Periodically, participants provided their subjective impression of the agent's character on a scale ranging from "nasty" to "nice", as well as their uncertainty of that impression. The two agents differed substantially in their level of harm aversion for others (κ_o): The "good" agent required more than five times the compensation per shock to the receiver than the "bad" agent. We tested the hypothesis that participants' predictions about the agents' choices reflected their estimates of the agents' harm aversion by fitting their trial-by-trial predictions with a computational model that combined our utility model of moral decision-making (Crockett et al., 2014) with a type of Bayesian learning model, the Hierarchical Gaussian Filter (HGF; Mathys, Daunizeau, Friston, & Stephan, 2011). The HGF provides a mathematical account of how people update beliefs about hidden states based on an

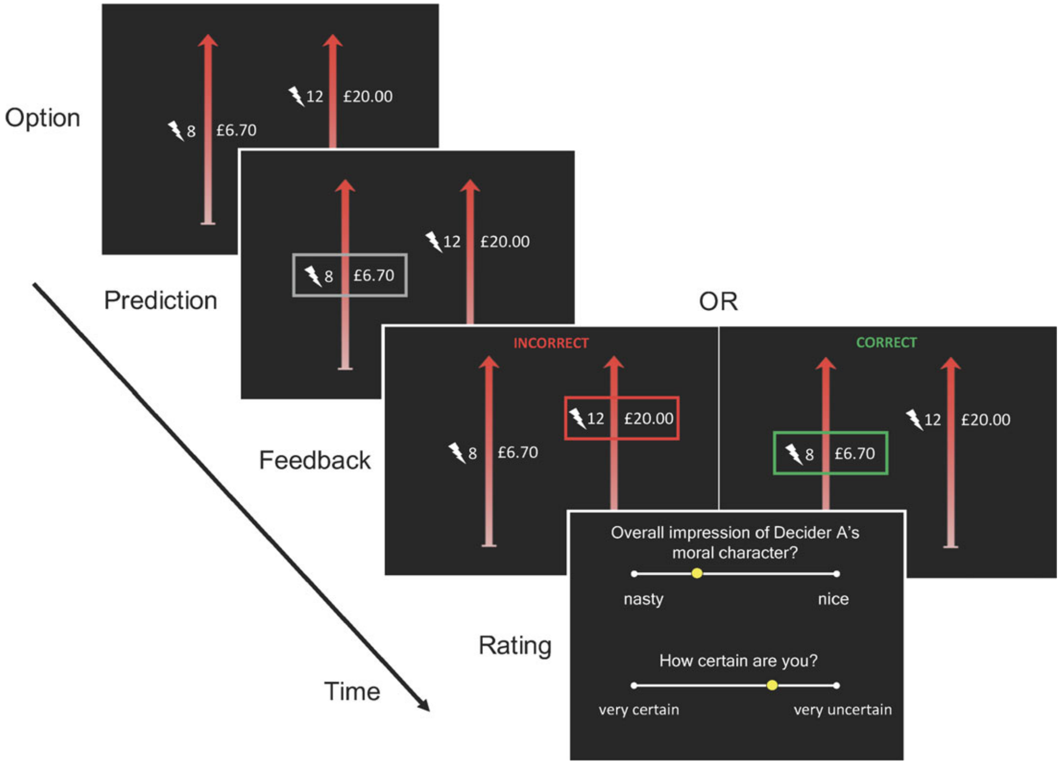

Fig. 5. Moral inference task. On every trial the agent chose between two options: more money for themselves plus more shocks for an anonymous receiver, or less money for themselves plus fewer shocks for the receiver. Participants predicted which option the agent would choose and subsequently received feedback about their accuracy. Every a few trials, participants provided their subjective impression of the agent's character on a scale ranging from "nasty" to "nice," as well as their uncertainty of that impression.

integration of prior beliefs about the hidden state and new observations. In our moral inference setting, beliefs are defined as probability distributions over a range of levels of harm aversion, and people update their beliefs about another agent's harm aversion based on observing that agent's decisions trading off money for themselves against pain for another. Our moral inference model specifies that participants make predictions about another agent's choices by passing their current belief about that agent's harm aversion (μ) through the utility model that describes choices in this same setting (Crockett et al., 2014; see Eq. 1a):

$$V_{\text{agent}}(\text{harm}) = (1-\mu)\Delta m - \mu\Delta s. \tag{3a}$$

$$P(\text{predict harm}) = \frac{1}{1+e^{-\beta V(\text{harm})}} \tag{3b}$$

β is a free parameter that describes how sensitive predictions are to the relative utility of different outcomes, or the prediction noise. When feedback is received (i.e., whether the prediction was correct or incorrect), beliefs are updated in proportion to the uncertainty of the belief, in accordance with Bayes' rule:

$$\mu_i \propto \mu_{i-1} + \sigma_{i-1} \cdot \delta \tag{4}$$

Conceptually, the belief about the agent's harm aversion at trial i (μ_i) is updated such that the updated belief is a function of the prior belief ($\mu_{i\text{-}1}$), plus a prediction error (δ) weighted by the uncertainty of the prior belief ($\sigma_{i\text{-}1}$). This update equation is similar in form to that in classical reinforcement learning models (e.g., Rescorla-Wagner models), but instead of multiplying the prediction error by a single learning rate, here the prediction error is weighted by the uncertainty of the prior belief, which changes dynamically across trials. In Bayesian inference, this uncertainty indicates how much learning still needs to occur. The model provides, for each participant, a trial-by-trial trajectory of belief estimates about each agent's harm aversion (μ); a trajectory of associated uncertainties on those beliefs (σ); and a global estimate of belief volatility (ω) that describes how flexibly participants update their beliefs about each agent's character based on observed decisions. The model explained significant variation in participants' predictions (capturing 87% of variance for lab studies, and 76% of variance for online studies) and outperformed a simpler Rescorla-Wagner learning model, which does not allow belief updates to vary in proportion to belief uncertainty.

Two notable findings emerge from the modeling results (Siegel et al., in press). First, the model demonstrates that similar computations are used to make moral decisions oneself, and to predict moral decisions in others. Second, by explicitly modeling the uncertainty and volatility of beliefs, we discovered that beliefs about the harm aversion of bad agents were more uncertain than beliefs about the harm aversion of good agents. As a result, beliefs about bad agents were more volatile than beliefs about good agents, indicating that in response to new information, people more readily revised their impressions about the bad agent than the good agent. These findings suggest a cognitive mechanism that could facilitate forgiveness, which involves changing one's attitudes toward transgressors (Beyens, Yu, Han, Zhang, & Zhou, 2015; Griswold, 2007; McCullough, Pargament, & Thoresen, 2001).

Control experiments showed that this asymmetry in belief updating was not observed when making inferences about low-skilled and high-skilled agents' competence. We adapted the inference paradigm such that participants predicted whether other agents would be able to score a certain number of basketball points in a given amount of time; competence was operationalized as the number of basketball points the agent could score per minute of game play. We found that a similar Bayesian updating model described predictions about basketball scoring as that which described predictions about moral choices, where beliefs about harm aversion were replaced with beliefs about competence. Notably, however, while participants formed more uncertain and volatile beliefs about bad agents' character, relative to good agents' character, beliefs about the competence of high-skill and low-skill agents were equally uncertain and volatile (Siegel et al., in press). This suggests that our experimental approach can not only be used to compare computational processes across different dimensions of moral cognition but can also shed light on differences between computational processes engaged in moral versus non-moral cognition.

3. Intersecting dimensions of moral cognition

Many important moral phenomena involve more than one dimension of moral cognition. The approach we describe here is most valuable in these cases, as it offers the possibility to quantify cross-dimension interactions hypothesized to underlie theoretically and practically interesting phenomena, such as person-centered morality, moral hypocrisy, and moral influence (Fig. 1). In this section, we will briefly discuss how the harm aversion paradigm and computational model outlined above could be utilized to investigate the neurocognitive basis of these phenomena.

3.1. *Person-centered morality: Inferences ∩ judgments*

Much research on moral judgment has focused predominantly on the evaluation of acts (i.e., act-centered), singling out features of acts that influence their moral evaluation (Baron, 2014; Malle et al., 2014; Shaver, 2012; Weiner, 1995), including the consequences of the act, the intentions of the actor, and the extent to which the actor caused the consequences (Cushman, 2008; Cushman, Murray, Gordon-McKeon, Wharton, & Greene, 2012b; Ginther et al., 2016; Karlovac & Darley, 1988; Shaver, 2012; Shultz & Wright, 1985; Shultz, Wright, & Schleifer, 1986; Weiner, 1995). However, in real-life situations, moral evaluations are often made with the knowledge of the moral character of the agent being evaluated. Recent work on such "person-centered" evaluations has explored how inferences about an agent's moral character influence moral judgments of that agent's acts (Alicke et al., 2015; Knobe, 2010; Tannenbaum et al., 2011; Uhlmann et al., 2015). Past research has demonstrated significant effects of character inferences on the evaluation of consequences, causation, and blame. We recently applied our computational framework to investigate how inferences about character affect the computation of blame via the evaluation of different aspects of moral acts, such as the consequences of the act and the degree to which the agent is causally responsible for those consequences.

Siegel, Crockett, and Dolan (2017) combined the moral judgment task (described in the previous section) with a manipulation of agents' moral character. Participants judged the blameworthiness/praiseworthiness of a series of decisions made by two agents with differing moral character. As in the moral inference task, we operationalized the moral character of the agents according to the harm aversion parameter in the decision model, where the "good" agent required more compensation to shock the receiver than the "bad" agent. To manipulate causation, the agent chose the more harmful option either passively (by default) or actively. To manipulate consequences, the agents' choices resulted in different amounts of profit for themselves and pain for the receiver. Here, because both the moral character of the agents (i.e., degree of harm aversion) and the features of their particular decision (i.e., amount of pain; and money) are quantitatively manipulated within the harm aversion framework, it is possible to demarcate and compare the contributions of person-level (i.e., moral character) and choice-level (i.e., money, pain, causation) features to moral judgments.

Results showed an effect of moral inference on the computation of blame: participants weighted the consequences of choices (i.e., profit and pain) more strongly in their judgments of bad agents' choices than good agents' choices. Specifically, profits mitigated the blameworthiness of harmful choices, and this effect was larger for the bad than the good agent. Meanwhile, blameworthiness scaled with the amount of pain inflicted, and this effect was also larger for the bad than the good agent. The increased weighting of consequences in judgments of bad agents may reflect enhanced attention toward the behaviors of potentially harmful individuals, the avoidance of whom may have benefits for survival (Tooby & Cosmides, 1992). Future work could usefully adapt Bayesian models of moral inference to interrogate how dynamically evolving impressions of agents' character shape subsequent moral judgments.

3.2. Moral hypocrisy: Decisions ∩ judgments

Moral hypocrisy occurs when people hold themselves to different moral standards than others, and it likely reflects a motivation to appear moral while behaving selfishly (Batson et al., 1997; Gino et al., 2016; Graham et al., 2015; Jordan, Sommers, Bloom, & Rand, 2017; Sharma et al., 2014; Szabados & Soifer, 2004). Researchers have operationalized hypocrisy in two complementary ways. One defines hypocrisy as a discrepancy between judgments and decisions: that is judging a decision to be wrong while nevertheless making that decision oneself (Batson & Thompson, 2001; Batson, Thompson, & Chen, 2002; Batson, Thompson, Seuferling, Whitney, & Strongman, 1999; Batson et al., 1997; Stone, Wiegand, Cooper, & Aronson, 1997). The second defines hypocrisy as a discrepancy between judgments of one's own and others' behavior, that is judging another person more harshly for transgressing than judging oneself for doing the same thing (Valdesolo & DeSteno, 2007, 2008). Moral hypocrisy is a widespread phenomenon, but its underlying mechanisms are not well understood.

Within our harm aversion framework, the first definition of moral hypocrisy arises at the intersection of moral decision-making and moral judgment (Fig. 1). Investigating hypocrisy within this framework can illuminate its underlying mechanisms. For example, participants can complete the moral decision task as well as the moral judgment task, and hypocrisy can be defined as a discrepancy between the indifference point for decisions and the indifference point for judgments. Given that moral decision-making involves a devaluation of ill-gotten gains via corticostriatal interaction (Crockett et al., 2017), one prediction is that hypocrites would show a reduced coupling between prefrontal regions that represent moral norms and striatal regions that represent the value of one's own actions. In other words, hypocrites may adequately represent moral norms but be unable to translate those norms into moral actions. This would be consistent with reports that criminal psychopaths, who have impaired corticostriatal function (Hosking et al., 2017), show intact moral judgments despite committing moral atrocities (Glenn, Raine, Schug, Young, & Hauser, 2009).

The second definition of hypocrisy can be operationalized within our framework as a discrepancy in the indifference point for judgment of one's own and others' decisions. One possible explanation for this kind of hypocrisy is readily apparent from our model of moral judgment (Eq. 2), where blame is mitigated by the profitability of harmful actions (Crockett et al., 2017; Siegel et al., 2017). It is well established that people value others' profits far less strongly than their own (Engel, 2011; Ruff & Fehr, 2014). Thus, it arises naturally from our model that people should blame others more than they blame themselves for the same profitable but harmful action, because profits for others are valued less (and thus would mitigate blame less) than profits for oneself. This account further predicts that those who place a higher value on others' rewards, that is those who are more generous, should be less hypocritical.

3.3. *Moral influence: Inferences $\cap$ decisions*

The question, "Can virtue be taught?" (Plato's *Meno*) has intrigued moral philosophers and educators for thousands of years. The effects of role models on prosocial and antisocial behaviors have been extensively studied since the early days of social psychology (Bandura, 1969; Cialdini & Goldstein, 2004; Hoffman, 1970; Macaulay & Berkowitz, 1970; Staub, 1971). For example, in one of Milgram's (1965) experiments, participants were instructed to deliver increasingly painful electric shocks to a receiver. The presence of confederates who refused to increase the shocks discouraged participants from escalating the pain (see also Rosenhan, 1969). More generally, research has shown that the degree of social influence is moderated by several factors, including how much people identify with and like the role model (Abrams & Hogg, 1990; Izuma & Adolphs, 2013), which may be moderated by inferences observers make about the role model's character. Thus, to better understand moral influence, we need a mechanistic account of how observers infer the moral character of role models and how such inferences might influence the observer's own decision-making.

Within our proposed framework (Fig. 1), moral influence can be understood as an interaction between moral inference and moral decision-making. By asking participants to complete the moral decision task before and after the moral inference task, it is possible to measure the effect that inferring the moral character of a role model has on one's own moral decisions. Additionally, we can probe how one's own baseline harm aversion moderates the extent of moral influence. In other words, is a morally "better" person (e.g., $\kappa_o = 0.7$) more susceptible to influence than a morally "worse" person (e.g., $\kappa_o = 0.3$)? Does the degree of influence depend on a person's objective or perceived similarity to the model (cf. Han, Kim, Jeong, & Cohen, 2017)? Our framework makes it possible to answer these questions because it offers a platform where the moral preferences of the role model and those of the participants can be precisely parameterized and compared quantitatively.

Understanding how role models influence the moral decisions of other people can help determine how to most effectively leverage the persuasive power of moral exemplars. For example, if the morally "worse" are more susceptible to influence, this would suggest sending role models to high-risk audiences, such as prisons; whereas if the morally "average" are more susceptible, this would suggest sending role models to more general audiences, such as schools. By jointly testing how the degree of influence depends on participants' actual similarity with the role model and perceived similarity with the role model, we can further identify channels for influence. For example, if perceived similarity between oneself and moral exemplars enhances susceptibility to influence, this suggests (perhaps counterintuitively) that making people feel they are more moral than they actually are, thus increasing perceived similarity with role models, could facilitate moral change.

4. Conclusion and future directions

In this contribution, we explored how investigating different dimensions of moral cognition (i.e., decision-making, judgment, inference) within the same experimental framework can facilitate a meaningful comparison of computational processes that may be common to multiple dimensions, and shed light on phenomena that emerge at the intersections of dimensions, such as moral hypocrisy, person-centered moral judgments, and moral influence. These phenomena are ubiquitous in everyday moral life (cf. Graham, 2014; Hofmann et al., 2014) but research on their cognitive and neural mechanisms is still in its infancy.

We used a recently developed experimental framework (Crockett et al., 2014, 2017; Siegel et al., 2017, in press) to illustrate how this approach can illuminate the cognitive mechanisms of harm-based moral cognition. We find preliminary support for our hypothesis that different dimensions of moral cognition are built upon basic utility computations that trade-off costs and benefits to oneself against costs and benefits to others. This generic "utility calculus" (see also Jara-Ettinger et al., 2016) accurately described moral decision-making, judgment and inference processes, as well as their underlying neural correlates, across several experiments in the domain of harm (Table 1). Individual differences in subcomponents of utility, namely harm aversion, were correlated across dimensions of moral cognition.

Although we have been focusing on harm-based morality throughout this paper, we believe that this approach can be applied to investigate other moral domains, such as trust, loyalty, or purity. For example, using this approach to investigate decisions, judgments, and inferences about trust would involve measuring decisions in the trust game (King-Casas et al., 2005; McCabe, Houser, Ryan, Smith, & Trouard, 2001; McCabe, Rigdon, & Smith, 2003) alongside moral judgments of others' decisions in the trust game and predictions about whether others are likely to make trustworthy decisions (e.g. Behrens et al., 2008; Diaconescu et al., 2014). Using this approach to compare computational processes across multiple moral domains may help resolve the debate about the centrality of harm in moral cognition (Schein & Gray, 2015, Schein & Gray, 2018). Finally, we note that similar computational frameworks can be used to model cognitive processes that fall outside the domain of morality entirely, as we have shown in our moral inference work comparing learning about morality versus. competence (Siegel et al., in press).

There are a few notable limitations to the approach we propose here. First, examining multiple dimensions of moral cognition within the same paradigm may *force*

Table 1
A common computational framework for measuring three dimensions of moral cognition

Dimension	Model
Moral decision-making	$V(\text{harm}) = (1 - k)\ \Delta m - k\ \Delta s$ $P(\text{choose harm}) = \frac{1}{1+e^{-\beta V(\text{harm})}}$
Moral judgment	Blame $= f\,(\Delta m, \Delta s, k)$[a]
Moral inference	$V_{\text{agent}}(\text{harm}) = (1 - \mu)\ \Delta m - \mu\ \Delta s$ $P(\text{predict harm}) = \frac{1}{1+e^{-\beta V(\text{harm})}}$ $\Delta\mu \propto \sigma\cdot\delta$

The decision model relates the subjective value of a harmful choice (V) to changes in money gained (Δm), changes in shocks delivered (Δs), and harm aversion (k) and transforms value differences into choices via a softmax function. In the judgment model, blame is a linear function of changes in money gained (Δm), changes in shocks delivered (Δs), and harm aversion (k). In the inference model, an agent's subjective value of a harmful choice (V) is simulated as a function of changes in money gained (Δm), changes in shocks delivered (Δs), and the current belief about the agent's harm aversion (μ); predictions about an agent's choice are based on a softmax transformation of simulated subjective value. Beliefs about harm aversion are updated proportionally to prediction errors (δ) weighted by the uncertainty of the prior belief (σ).

[a] The blame model is a linear combination of Δm, Δs, k_o, k_s and their interactions. For details, please see Section 2.2.

computational processes to be shared across dimensions, when in reality different dimensions might employ rather different computations. Second, building paradigms that are amenable to computational modeling can require trading off real-world richness for methodological rigor. Finally, identifying a computational model that provides a good fit to behavior or brain activity does not guarantee that the identified model is the *best* or most accurate model (Mars, Shea, Kolling, & Rushworth, 2012); an important and often overlooked aspect of computational cognitive modeling is falsifying candidate models in light of observed data (Palmintieri et al. 2017).

In spite of these limitations, computational frameworks may be especially useful in providing quantitative measures of individual differences in moral cognition that do not rely on self-report, which can be less reliable in measuring traits that have a strong social desirability component. Such individual differences are also likely to be meaningful in the context of psychiatric disorders which often involve social difficulties (Mendez, 2009), providing biomarkers for intact and affected moral cognition complementary to traditional diagnosis. The model parameters could serve as an intermediate level (or “cognitive phenotype”; cf. Montague, Dolan, Friston, & Dayan, 2012) between biological and phenomenological descriptions of how a given (sub-)clinical and psychiatric condition influences moral cognition and behavior, such as obsessive-compulsive disorder (OCD; Harrison et al., 2012), psychopathy (Blair, 2007; Marsh et al., 2011), and personality disorder (Tyrer, Reed, & Crawford, 2015). This approach thus holds great promise not just for advancing our understanding of human morality, but also for reducing human suffering in health and disease.

Acknowledgments

The authors thank the two anonymous reviewers for their useful comments and suggestions. H.Y. was supported by British Academy Newton International Fellowship (NF160700). J.Z.S. was supported by a Wellcome Trust Society and Ethics studentship (104980/Z/14/Z). This work was supported by a Wellcome Trust ISSF Award (204826/Z/16/Z), the John Fell Fund, the Academy of Medical Sciences (SBF001\1008), and the John Templeton Foundation Beacons Project. The funders did not play a role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of this manuscript.

References

Abrams, D., & Hogg, M. A. (1990). Social identification, self-categorization and social influence. *European Review of Social Psychology*, *1*(1), 195–228.

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation a historical tour. *Perspectives on Psychological Science*, *10*(6), 790–812.

Alicke, M. D., & Zell, E. (2009). Social attractiveness and blame. *Journal of Applied Social Psychology*, *39* (9), 2089–2105.

Bandura, A. (1969). *Principles of behavior modification* (p. 1969). New York: Holt, Rinehart & Winston.

Baron, J. (2014). Moral Judgment. In E. Zamir, & D. Teichman (Eds.), *The Oxford handbook of behavioral economics and the law* (pp. 61–89). New York: Oxford University Press.

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral Judgment and Decision Making. In G. Keren, & C. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 478–515). Wiley Blackwell, West Sussex: UK.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, *40*(2), 290.

Batson, C. D., Fultz, J., & Schoenrade, P. A. (1987). Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of Personality*, *55*(1), 19–39.

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology*, *72*(6), 1335.

Batson, C. D., & Thompson, E. R. (2001). Why don't moral people act morally? Motivational considerations. *Current Directions in Psychological Science*, *10*(2), 54–57.

Batson, C. D., Thompson, E. R., & Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, *83*(2), 330.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, *77*(3), 525.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243.

Behrens, T. E., Hunt, L. T., & Rushworth, M. F. (2009). The computation of social behavior. *Science*, *324* (5931), 1160–1164.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249.

Beyens, U., Yu, H., Han, T., Zhang, L., & Zhou, X. (2015). The strength of a remorseful heart: Psychological and neural basis of how apology emolliates reactive aggression and promotes forgiveness. *Frontiers in psychology*, *6*, 1611.

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences*, *11*(9), 387–392.

Bostyn, D. H., & Roets, A. (2017). Trust, trolleys and social dilemmas: A replication study. *Journal of Experimental Psychology: General*, *146*(5), e1–e7.

Buckholtz, J. W. (2015). Social norms, self-control, and the value of antisocial behavior. *Current Opinion in Behavioral Sciences*, *3*, 122–129.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, *70*(3), 560–572.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Crockett, M. J. (2016). How formal models can illuminate mechanisms of moral judgment and decision making. *Current Directions in Psychological Science*, *25*(2), 85–90.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, *107*(40), 17433–17438.

Crockett, M. 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.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., Grosse-Rueskamp, J. M., Dayan, P., & Dolan, R. J. (2015). Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. *Current Biology*, *25*(14), 1852–1859.

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

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012a). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2.

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2012b). Judgment before principle: Engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience*, *7*(8), 888–895.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In E. A. Phelps, T. W. Robbins, & M. Delgado (Eds.), *Affect, learning and decision making, attention and performance XXIII*. New York, NY: Oxford University Press.

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., & Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology*, *10*(9), e1003810.

Diaconescu, A. O., Mathys, C., Weber, L. A., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618–634.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583–610.

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

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

Fehr, E., & Krajbich, I. (2014). Social preferences and the brain. *Neuroeconomics* (2nd ed.) (pp. 193–218). London: Academic Press.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441.

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *NeuroImage*, *105*, 347–356.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Freud, S. (1961). *Civilization and its discontents*. New York, NY: Norton.

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., & Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous-and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences*, 201802523.

Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*, *19*(12), 1727.

Gert, B. (2004). *Common morality: Deciding what to do*. New York: Oxford University Press.

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond purity: Moral disgust toward bad character. *Psychological Science*, *28*(1), 80–91.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological science*, *20*(3), 393–398.

Gino, F., Norton, M. I., & Weber, R. A. (2016). Motivated Bayesians: Feeling moral while acting egoistically. *The Journal of Economic Perspectives*, *30*(3), 189–212.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *Journal of Neuroscience*, *36*(36), 9420–9434.

Glenn, A. L., Raine, A., Schug, R. A., Young, L., & Hauser, M. (2009). Increased DLPFC activity during moral decision-making in psychopathy. *Molecular Psychiatry*, *14*(10), 909–911.

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.

Graham, J. (2014). Morality beyond the lab. *Science*, *345*(6202), 1242–1242.

Graham, J., Meindl, P., Koleva, S., Iyer, R., & Johnson, K. M. (2015). When values and behavior conflict: Moral pluralism and intrapersonal moral hypocrisy. *Social and Personality Psychology Compass*, *9*(3), 158–170.

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

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.

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.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, *106*(30), 12506–12511.

Griswold, C. (2007). *Forgiveness: A philosophical exploration*. Cambridge University Press.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235.

Han, H., Kim, J., Jeong, C., & Cohen, G. L. (2017). Attainable and relevant moral exemplars are more effective than extraordinary exemplars in promoting voluntary service engagement. *Frontiers in Psychology*, *8*, 283.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, *324*(5927), 646–648.

Harrison, B. J., Pujol, J., Soriano-Mas, C., Hernández-Ribas, R., López-Solà, M., Ortiz, H., Alonso, P., Deus, J., Menchon, J. M., Real, E., Segalàs, C., Contreras-Rodríguez, O., Blanco-Hinojo, L., & Cardoner, N. (2012). Neural correlates of moral sensitivity in obsessive-compulsive disorder. *Archives of General Psychiatry*, *69*(7), 741–749.

Helzer, E. G., & Critcher, C. R. (2018). What do we evaluate when we evaluate moral character? In K. Gray, & J. Graham (Eds.), *Atlas of moral psychology* (pp. 99–107). New York, NY: The Guilford Press.

Hoffman, M. L. (1970). Moral development. In P. H. Mussen (Ed.), *Handbook of child psychology*. Vol. *2* (pp. 261–293). New York, NY: Wiley.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345* (6202), 1340–1343.

Hosking, J. G., Kastman, E. K., Dorfman, H. M., Samanez-Larkin, G. R., Baskin-Sommers, A., Kiehl, K. A., Newman, J. P., & Buckholtz, J. W. (2017). Disrupted prefrontal regulation of striatal subjective value signals in psychopathy. *Neuron*, *95*(1), 221–231.

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*(5879), 1092–1095.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*(2), 451–462.

Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, *78*(3), 563–573.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356–368.

Karlovac, M., & Darley, J. M. (1988). Attribution of responsibility for accidents: A negligence law analogy. *Social Cognition*, *6*(4), 287–318.

Keane, W. (2015). *Ethical life: Its natural and social histories*. Princeton, NJ: Princeton University Press.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83.

Kitcher, P. (1993). The evolution of human altruism. *The Journal of Philosophy*, *90*(10), 497–516.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957.

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

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908.

Konovalov, A., Hu, J., & Ruff, C. C. (2018). Neurocomputational approaches to social behavior. *Current Opinion in Psychology*, *24*, 41–47.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist Bias. *Neuron*, *93*(3), 480–490.

Lewis, H. B. (1971). Shame and guilt in neurosis. *Psychoanalytic Review*, *58*(3), 419.

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science*, *7*(2), 230–242.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Macaulay, J., & Berkowitz, L. (1970). *Altruism and helping behavior*. New York, NY: Academic Press.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly Journal of Experimental Psychology*, *65*(2), 252–267.

Marsh, A. A., Finger, E. C., Fowler, K. A., Jurkowitz, I. T., Schechter, J. C., Yu, Pine, Pine, D. S., & Blair, R. J. R. (2011). Reduced amygdala–orbitofrontal connectivity during moral judgments in youths with disruptive behavior disorders and psychopathic traits. *Psychiatry Research: Neuroimaging*, *194*(3), 279–286.

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.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835.

McCabe, K., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, *52*(2), 267–275.

McCullough, M. E., Pargament, K. I., & Thoresen, C. E. (Eds.). (2001). *Forgiveness: Theory, research, and practice*. Guilford Press.

Mendez, M. F. (2009). The neurobiology of moral behavior: Review and neuropsychiatric implications. *CNS Spectrums*, *14*(11), 608–620.

Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Human Relations*, *18*(1), 57–76.

Mill, J. S. (1863/1998). *Utilitarianism*. Oxford: Oxford University Press.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80.

Nadler, J. (2012). Blaming as a social process: The influence of character and moral emotion on blame. *Law and Contemporary Problems*, *75*, 1.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*(1), 35–53.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences*, *21*(6), 425–433.

Planas, R. (2012). Victoria Soto, Newtown teacher, emerges as hero after shooting. Available at: https://www.huffingtonpost.com/. Accessed May 19, 2018.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489* (7416), 427–430.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making*, *3*(4), 263–277.

Rosenhan, D. L. (1969). The kindnesses of children. *Young Children*, *25*(1), 30–44.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549–562.

Sáez, I., Zhu, L., Set, E., Kayser, A., & Hsu, M. (2015). Dopamine modulates egalitarian behavior in humans. *Current Biology*, *25*(7), 912–919.

Schein, C., & Gray, K. (2014). The prototype model of blame: Freeing moral cognition from linearity and little boxes. *Psychological Inquiry*, *25*(2), 236–240.

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, *41*(8), 1147–1163.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. (2015). Optimal inference with suboptimal models: Addiction and active Bayesian inference. *Medical Hypotheses*, *84*(2), 109–117.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science*, *24*(2), 125–130.

Sharma, E., Mazar, N., Alter, A. L., & Ariely, D. (2014). Financial deprivation selectively shifts moral standards and compromises moral decisions. *Organizational Behavior and Human Decision Processes*, *123*(2), 90–100.

Shaver, K. (2012). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, NY: Springer, New York.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667–677.

Shultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *17*(2), 97–108.

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development*, *57*(1), 177–184.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211.

Siegel, J. Z., Mathys, C., Rutledge, R., & Crockett, M. J. (in press). Beliefs about bad people are volatile. *Nature Human Behaviour*, https://doi.org/10.1038/s41562-018-0425-1.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157–1162.

Smith, A. (1759/2002). *The theory of moral sentiments*. Cambridge, UK: Cambridge University Press.

Staub, E. (1971). Helping a person in distress: The influence of implicit and explicit "rules" of conduct on children and adults. *Journal of Personality and Social Psychology*, *17*(2), 137.

Stone, J., Wiegand, A. W., Cooper, J., & Aronson, E. (1997). When exemplification fails: Hypocrisy and the motive for self-integrity. *Journal of Personality and Social Psychology*, *72*(1), 54.

Szabados, B., & Soifer, E. (2004). *Hypocrisy: Ethical investigations*. Orchard Park, NY: Broadview Press.

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

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York, NY: Oxford University Press.

Tyrer, P., Reed, G. M., & Crawford, M. J. (2015). Classification, assessment, prevalence, and effect of personality disorder. *The Lancet*, *385*(9969), 717–726.

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

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy social groups and the flexibility of virtue. *Psychological Science*, *18*(8), 689–690.

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology*, *44*(5), 1334–1338.

Volz, L. J., Welborn, B. L., Gobel, M. S., Gazzaniga, M. S., & Grafton, S. T. (2017). Harm to self outweighs benefit to others in moral decision making. *Proceedings of the National Academy of Sciences*, *114*(30), 7963–7968.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct* (Vol. xvi). New York, NY: Guilford Press.

Wojciszke, B., Parzuchowski, M., & Bocian, K. (2015). Moral judgments and impressions. *Current Opinion in Psychology*, *6*, 50–54.

Xie, W., Yu, B., Zhou, X., Sedikides, C., & Vohs, K. D. (2014). Money, moral transgressions, and blame. *Journal of Consumer Psychology*, *24*(3), 299–306.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *Neuroimage*, *40*(4), 1912–1920.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, *107*(15), 6753–6758.

Yu, H., Gao, X., Zhou, Y., & Zhou, X. (2018). Decomposing gratitude: Representation and integration of cognitive antecedents of gratitude in the brain. *Journal of Neuroscience*, *38*, 4886–4898.

Yu, H., Hu, J., Hu, L., & Zhou, X. (2014). The voice of conscience: Neural bases of interpersonal guilt and compensation. *Social Cognitive and Affective Neuroscience*, *9*(8), 1150–1158.

Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., ... & Hsu M. (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience*, *17* (10), 1319.