**Computational modeling of moral decisions**

Molly J. Crockett, Department of Experimental Psychology, University of Oxford

## Abstract

The cognitive processes that give rise to moral decisions have long been the focus of intense study. Here I illustrate how computational approaches to studying moral decision-making can advance this endeavor. Computational methods have traditionally been employed in the domains of perceptual and reward-based learning and decision-making, but until recently had not been applied to the study of moral cognition. Using examples from recent studies I show how computational properties of choices provide can provide novel insights into moral decision-making. I conclude with an exploration of new research avenues that arise from these insights, such as how uncertainty in choice shapes morality, and how moral decision-making can be viewed as a learning process.

## Introduction

Moral decisions often involve tradeoffs between personal benefits and preventing harm to others. How do we decide when faced with such dilemmas? And how do we judge the moral decisions of others? These questions have long been the focus of intense study in philosophy, psychology, and more recently neuroscience. To investigate these questions, researchers have employed a variety of methods, ranging from hypothetical thought experiments (Cushman, Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Jonathan Haidt, 2001) to virtual reality environments (David, McDonald, Mott, & Asher, 2012; Slater et al., 2006) to real moral decisions (Batson, Duncan,

Ackerman, Buckley, & Birch, 1981; Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; FeldmanHall, Mobbs, et al., 2012; Hein, Silani, Preuschoff, Batson, & Singer, 2010; Valdesolo & DeSteno, 2008). In this chapter I will review recent work illustrating new approaches to investigating moral cognition that borrow from methods traditionally employed in the domains of perceptual and reward-based decision-making. Throughout, I will focus on moral cognition concerned with harm and care towards others.[1]

Early studies in this area examined the extent to which people would invest effort in helping others in need, and how features of social situations influenced helping behavior. These experiments often involved elaborately staged situations, for example with confederates trapped by falling bookcases (Ashton & Severy, 1976), having epileptic seizures (Darley & Latane, 1968), or collapsing in subway cars (J. A. Piliavin & Piliavin, 1972). In classic studies by Batson et al. (1981) subjects were given the opportunity to reduce the number of electric shocks delivered to a confederate by taking on some of the shocks themselves. These studies laid the groundwork for much of what we know about altruism and moral behavior and have high ecological validity. However, because these procedures generally gather only a single data point per subject, they provide rather sparse data sets that do not allow for interrogation of the computations underlying choices. These methods are also impractical for investigating the neural mechanisms of decision-making.

[1] Although there is ample evidence suggesting morality is about more than harm and care (Graham et al., 2011; J. Haidt, 2007), I focus on this domain here because the bulk of research on moral cognition has investigated this particular aspect of morality, and because harm and care towards others has obvious parallels with behavioural economic studies of social preferences and work in computational neuroscience about the valuation of outcomes.

Perhaps the most widely used method for studying moral cognition is examining how people respond to hypothetical scenarios. For example, in the classic "trolley problem" (Foot, 1967; Thomson, 1976), a trolley is hurtling out of control down the tracks toward five workers, who will die if you do nothing. You and a large man are standing on a footbridge above the tracks. You realize that you can push the large man off the footbridge onto the tracks, where his body will stop the trolley and prevent it from killing the workers. Is it morally permissible to push the man, killing him but saving the five workers? By systematically varying features of these scenarios, researchers have uncovered a trove of insights into the mechanics of moral judgment, illuminating important influences of factors such as intentionality (L. Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; Liane Young, Cushman, Hauser, & Saxe, 2007), actions (Cushman, 2013; Cushman et al., 2006; Spranca, Minsk, & Baron, 1991), physical contact (Cushman et al., 2006; Greene et al., 2009), and incidental emotions (Eskine, Kacinik, & Prinz, 2011; Horberg, Oveis, & Keltner, 2011; Ugazio, Lamm, & Singer, 2012), among others.

However, hypothetical scenarios may be less useful for investigating moral *decisions*, as it is unclear to what extent judgments in these scenarios reflect how people would actually behave when faced with a real moral decision. This question was addressed directly in a recent study (FeldmanHall, Mobbs, et al., 2012). Subjects were given the opportunity to spend up to £1 to reduce the intensity of an electric shock delivered to a confederate sitting in the next room, whom they had recently met. Decisions were probed in two conditions. In the "real" condition, subjects were led to believe they would be making decisions with real consequences for themselves and the

confederate. In the "hypothetical" condition, subjects were explicitly instructed that their decisions were hypothetical and would not have consequences for themselves or the confederate. Subjects behaved differently in the real versus hypothetical conditions, and real versus hypothetical decisions engaged overlapping but distinct neural networks (FeldmanHall, Dalgleish, et al., 2012). Another similarly motivated recent study showed that decisions about whether to cooperate with an anonymous other differed in real versus hypothetical situations (Vlaev, 2012). Collectively this work suggests that moral decisions as probed by hypothetical scenarios may not necessarily be reflective of true moral preferences.

How, then, might we investigate the psychological (and neural) processes governing moral decisions? Behavioral economic games offer a tool for probing social preferences by measuring how people make decisions that have real monetary consequences for themselves and anonymous others, as well as the neural processes underlying such decisions (Camerer, 2003; Glimcher & Fehr, 2013). For instance, in the dictator game, participants are given some money and can share none, some or all of it with an anonymous other person. The amount shared is reflective of the value people place on rewards to others, as well as attitudes toward inequality: the more people value rewards to others, and the more they dislike being in an advantageous position relative to someone else, the more money they will transfer to the other person (Camerer, 2003; Glimcher & Fehr, 2013).

There are several features of economic games that make them well suited for probing moral decision-making. Because they are incentivized, choices in these paradigms are faithful reflections of people's actual preferences. This is especially critical

in the case of moral decision-making. Self-report questionnaires aimed at measuring moral preferences suffer from the obvious limitation that social desirability is likely to have a strong influence on people's answers. When there is no cost to answering "no" to the question of whether you would harm someone else for personal gain, most people would do so to preserve their reputation, regardless of their actual preferences. Subject anonymity is important for similar reasons. If subjects interact face-to-face with one another, then prosocial decisions could be reflective of people's selfish desire to preserve their own reputation, rather than their true preferences with regards to the welfare of others.

Perhaps even more importantly, economic games are also amenable to building computational models of choice processes and linking these models to neural activity. Despite progress in mapping the facets of moral cognition, still very little is known about the computational mechanisms that underlie moral decisions and indeed social cognition more broadly (Korman, Voiklis, & Malle, 2015). Formalizing the components of cognitive processes and how these components interact using a model-based approach has advanced our understanding of many other aspects of cognition, including perception, reasoning, learning, language, and reward-based decision-making. Applying a similar model-based approach to moral cognition will yield similar progress and generate novel predictions about the nature of moral decision-making and its neural basis.

Decades of research on social preferences using economic games has demonstrated that when it comes to monetary exchanges, people do value others' outcomes to a certain extent – although they care about their own outcomes far more (Charness & Rabin, 2002; Engel, 2011; Fehr & Schmidt, 1999). This work provides

proof-of-principle that even complex social behaviors can be accurately described using formal mathematical models. However, it is unclear to what extent these paradigms probe *moral* preferences. Gray, Young and Waytz (2012) argue that the essence of a moral transgression is an intentional agent causing harm to a suffering moral patient (Gray, Waytz, & Young, 2012; Gray, Young, & Waytz, 2012). Although economic games certainly capture intentional decisions, whether they induce suffering is debatable. Given that the worst possible outcome for a recipient in a dictator game is to receive nothing, and even putatively "unfair" transfers in the dictator game (i.e., < 50%) yield benefits for the recipient, it seems inappropriate to construe the dictator game as a moral decision.

Computing the costs of others' suffering is central to the process of making moral decisions. Although the bulk of research on value-based decision-making has investigated decisions involving only oneself, several studies have examined the neural basis of decisions that affect others. There is growing evidence that computing the value of outcomes to others engages neural mechanisms similar to those used to compute the value of one's own outcomes. At the heart of this process is a value-based decision-making circuitry that includes the striatum and the ventromedial prefrontal cortex (vmPFC). The current consensus is that the vmPFC computes the subjective value of the chosen option when a choice is made, as well as the experienced value of the outcome when it is received (Clithero & Rangel, 2013). Meanwhile, the striatum computes value differences between expectations and experiences, i.e., *prediction errors* (Clithero & Rangel, 2013). Decisions affecting others engage the striatum and vmPFC in a similar manner to decisions that affect only oneself (Fehr & Krajbich, 2013). For example, choosing to donate money to anonymous others or charities activates the vmPFC and

striatum in a similar manner to choices that reap rewards for oneself (Harbaugh, Mayr, & Burghart, 2007; Hare, Camerer, Knoepfle, O'Doherty, & Rangel, 2010; Zaki & Mitchell, 2011). A recent meta-analysis comparing the neural correlates of rewards to self and rewards to others (i.e., vicarious rewards showed that self and vicarious rewards engage overlapping regions of vmPFC (Morelli et al., 2015).

Thus far the majority of studies investigating social preferences have examined decisions involving rewarding outcomes to others. How people value aversive outcomes to others is less well understood. Recently my colleagues and I developed new methods for measuring how people compute the value of painful outcomes to others, relative to themselves (Crockett et al., 2014). In the following sections I will describe these methods and the questions that have arisen out of studies employing it.

**Quantifying the costs of harm to self and others**

We are able to quantify how much people value harm to self versus others by inviting them to trade of profits for themselves against pain to either themselves or others. In essence this involves measuring how much people are willing to pay to prevent pain to themselves and others, as well as how much compensation people require to increase pain to themselves and others. By combining questions such as these with computational models of choice, we are able to extract the precise values people ascribe to their own negative outcomes as well as those of others.

Two participants visit the lab in each experimental session. They arrive at staggered times and are led to different rooms to ensure they do not see or interact with one another. Next, each participant is led through a well-validated pain thresholding

procedure in which we use an electric stimulation device (Digitimer DS5) to deliver electric shocks to the left wrist of our volunteers (Story et al., 2013; Vlaev, Seymour, Dolan, & Chater, 2009). Shocks delivered by this device can range from imperceptible to intolerably painful, depending on the electric current level. Importantly, the shocks are safe and don't cause any damage to the skin.

In the thresholding procedure, we start by delivering a shock at a very low current level – 0.1 milliamps (mA) – that is almost imperceptible. We then gradually increase the current level, shock by shock, and the volunteer rates each shock on a scale from 0 (imperceptible) to 10 (intolerable). We stop increasing the current once the volunteer's rating reaches a 10. For the shocks used in the experiment we use a current level that corresponds to a rating of 8 out of 10, so the shocks are unpleasant, but not intolerable.

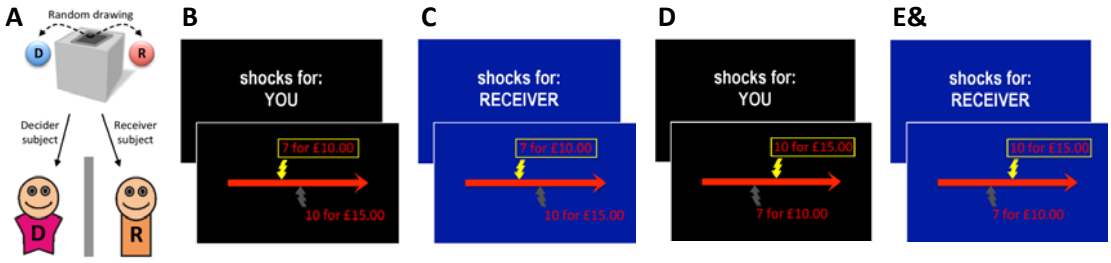
Critically, this procedure allows us to ensure that (a) the stimuli delivered in our experiment are in fact painful, (b) the stimuli are subjectively matched for the two participants, which is a necessary for comparing the valuation of pain to self and others, and (c) subjects experience the stimuli about which they will later be making decisions, which is important for minimizing ambiguity in those decisions.

Next, the participants are randomly assigned to the roles of "decider" and "receiver". We used a randomization procedure that preserved subjects' anonymity, while at the same time confirmed the existence of another participant in the experiment and transparently provided a fair allocation of roles. Anonymity here is essential because we want to isolate the contribution of moral *preferences* for avoiding harm to others, independently from the influence of selfish concerns about preserving one's own reputation and avoiding retaliation, both of which could readily explain altruistic

behavior in the context of a face-to-face interaction where identities are common knowledge (Fehr & Krajbich, 2013). In addition, anonymity is important for establishing a baseline level of moral preferences. It is well known that characteristics of the victim influence helping behavior (Penner, Dovidio, Piliavin, & Schroeder, 2005; I. M. Piliavin, Piliavin, & Rodin, 1975; Stürmer, Snyder, Kropp, & Siem, 2006), but the influences of these factors can only be documented relative to baseline (Fehr & Krajbich, 2013).

Following this the decider completes a decision task (Fig. 1). In this task they make a series of approximately 160 decisions involving tradeoffs between profits for themselves against pain for either themselves or the receiver. In each trial deciders choose between less money and fewer shocks, vs. more money and more shocks. The money is always for the decider, but in half the trials the shocks are for the decider (Fig 1A and 1C) and in the other half the shocks are for the receiver (Fig 1B and 1D). In all trials, if the decider fails to press a key within 6 s the highlighted default (top) option is registered; if the decider presses the key, the alternative (bottom) option is highlighted and registered instead. In half the trials, the alternative option contains more money and shocks than the default (Fig 1A and 1B), and in the other half the alternative option contains less money and fewer shocks than the default (Fig 1C and 1D). To avoid habituation and preserve choice independence no money or shocks are delivered during the task. Instead, one trial is selected by the computer and implemented at the end of the experiment, and subjects are made aware of this. Subjects are also instructed that their decisions will be completely anonymous and confidential with respect to both the receiver and the experimenters.

**Figure 1. A paradigm for extracting the subjective value of harm to self and others.** (A) Subjects remained in separate testing rooms at all times and were randomly assigned to roles of decider and receiver. (B–E) In each trial the decider chose between less money and fewer shocks, vs. more money and more shocks. The money was always for the decider, but in half the trials the shocks were for the decider (B and D) and in the other half the shocks were for the receiver (C and E). In all trials, if the decider failed to press a key within 6 s the highlighted default (top) option was registered; if the decider pressed the key, the alternative (bottom) option was highlighted and registered instead. In half the trials, the alternative option contained more money and shocks than the default (B and C), and in the other half the alternative option contained less money and fewer shocks than the default (D and E). Adapted from (Crockett et al., 2014).




The key dependent measure that can be extracted from this paradigm is a pair of subject-specific *harm aversion* parameters that are derived from a computational model of subjects' choices. These parameters capture the subjective costs of harm to self and others and are proportional to the amount of money subjects are willing to pay to prevent an additional shock to themselves and the receiver; in other words, harm aversion represents an "exchange rate" between money and pain. When we began this research we

had very little basis for predicting what these exchange rates would look like. So in our first study we used a staircasing procedure that homes in on subjects' exchange rates by estimating the exchange rate after each choice and then generating subsequent choices that will provide the most new information about the exchange rates. One obvious drawback of this approach is that subjects' preferences influence the choice set they see in the task, and if this is discovered there is the possibility that subjects could consciously "game" the task. There is also the issue that the context in which choices are made can influence the choices themselves; for example, people are willing to pay more to avoid a medium-intensity shock when it is presented alongside low-intensity shocks than when it is presented in the context of high-intensity shocks (Vlaev et al., 2009). Thus it is preferable to present all subjects with the same set of choices that are pre-determined to be able to detect exchange rates within the range expected in the population. We did this in our second study once having determined the range of exchange rates expected in the population which were recovered using the staircasing procedure described above.

One of the most common methods for modeling decision-making involves two steps (Daw, 2011). In the first step, a *value model* is specified that relates features of the choice options to their underlying subjective values. For instance, a very basic value model for a dictator game might simply state that the subjective value of a given choice in the dictator game consists of the amount of money kept for oneself, multiplied by a self-weight parameter that indicates how much one cares for their own outcome, plus the amount of money transferred, multiplied by an other-weight parameter that indicates how much one cares for the others' outcome. In the second step, a *choice model* is specified that passes the subjective values (coming from the value model) through a stochastic

decision process whereby choice options with higher subjective values are more likely to be selected. Critically, there is noise in the selection process, and the amount of noise is modulated by additional parameters. Here I will describe the findings that arise from the value model; in the next section I will discuss the choice model.

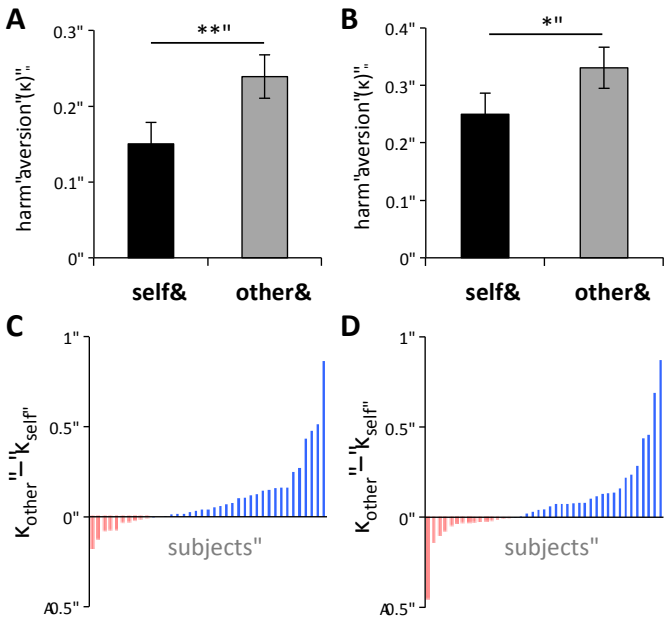
In our studies we fit a series of value models to subjects' choices and used Bayesian model comparison to identify the one that explained their choices the best (Burnham & Anderson, 2002). The best value model turned out to be quite simple. Essentially, the value model indicates that differences in subjective value between the choice options depend on the following parameters:

- A *self harm aversion* parameter that captures how much one prefers to avoid shocks to oneself
- An *other harm aversion* parameter that captures how much one prefers to avoid shocks to others
- A *loss aversion* parameter that weights negative outcomes (i.e., monetary losses and increases in shocks) more strongly than positive ones (i.e., monetary gains and decreases in shocks)

Strikingly, when we examined these estimates we found in both studies that harm aversion for others was greater on average than harm aversion for self (Fig 2A,B). In other words, people were willing to pay more to prevent shocks to others than the same shocks to themselves, and likewise they required greater compensation to increase shocks to others than to increase shocks to themselves. This "hyperaltruistic" disposition was present in the majority of subjects (Fig 2C,D). Notably, hyperaltruism is not predicted by existing economic models of social preferences (Charness & Rabin, 2002; Fehr &

Schmidt, 1999) and even more recent research linking empathy and altruism would not predict that people would care about avoiding others' pain *more* than their own (Batson et al., 1981; Bernhardt & Singer, 2012; Hein et al., 2010).

**Figure 2. Harm to others outweighs harm to self in moral decision-making.** (A-B) Estimates of harm aversion for self and other in study 1 (A) and 2 (B). Error bars represent SEM difference between κself and κother. (C-D) Distribution of hyper- altruism (κother − κself) across subjects in study 1 (C) and 2 (D). *P < 0.05, **P < 0.01.




What, then, could explain hyperaltruistic harm aversion? We suggested two potential explanations that are not mutually exclusive (Crockett et al., 2014). First, harming others carries a cost of moral responsibility that harming oneself does not, and this cost could explain why people are willing to pay more to avoid harming others than themselves. This account gels with work showing that in hypothetical scenarios people dislike being responsible for bad outcomes (Leonhardt, Keller, & Pechmann, 2011).

The second explanation stems from the fact that decisions affecting other people are necessarily uncertain because we can never truly know what another person's subjective experience is like (Harsanyi, 1955; Nagel, 1974). In the case of pain, there is a risk that what is tolerably painful for oneself might be intolerably painful for another. Because we want to avoid imposing intolerable costs on another person, we may adopt a risk-averse choice strategy, erring on the side of caution when it comes to actions that could potentially harm others. Indeed, many of our subjects expressed this logic when explaining their choices post-hoc. One typical subject reported, "I knew what I could handle, but I wasn't sure about the other person and didn't want to be cruel." In this way, uncertainty about others' subjective experience in the presence of social norms that strongly proscribe harming others could naturally lead to the pattern of hyperaltruistic choices that we observe. Intriguingly, empirical support for the uncertainty explanation comes directly from the choice model.

**The role of uncertainty in moral decisions**

As described above, the choice model translates information about the subjective value of choice options into actual decisions. Typically this model takes the form of a softmax equation (Daw, 2011). Our choice model contained two parameters – a *choice accuracy* parameter[2] and an *irreducible noise* parameter -- that respectively capture the noisiness of "difficult" choices (where the choice options are similarly attractive) and "easy" choices (where one of the choice options is substantially more attractive than the other). When choice accuracy is high, the more highly valued option will be deterministically chosen

[2] This is sometimes referred to as the *inverse temperature* or *softmax slope*

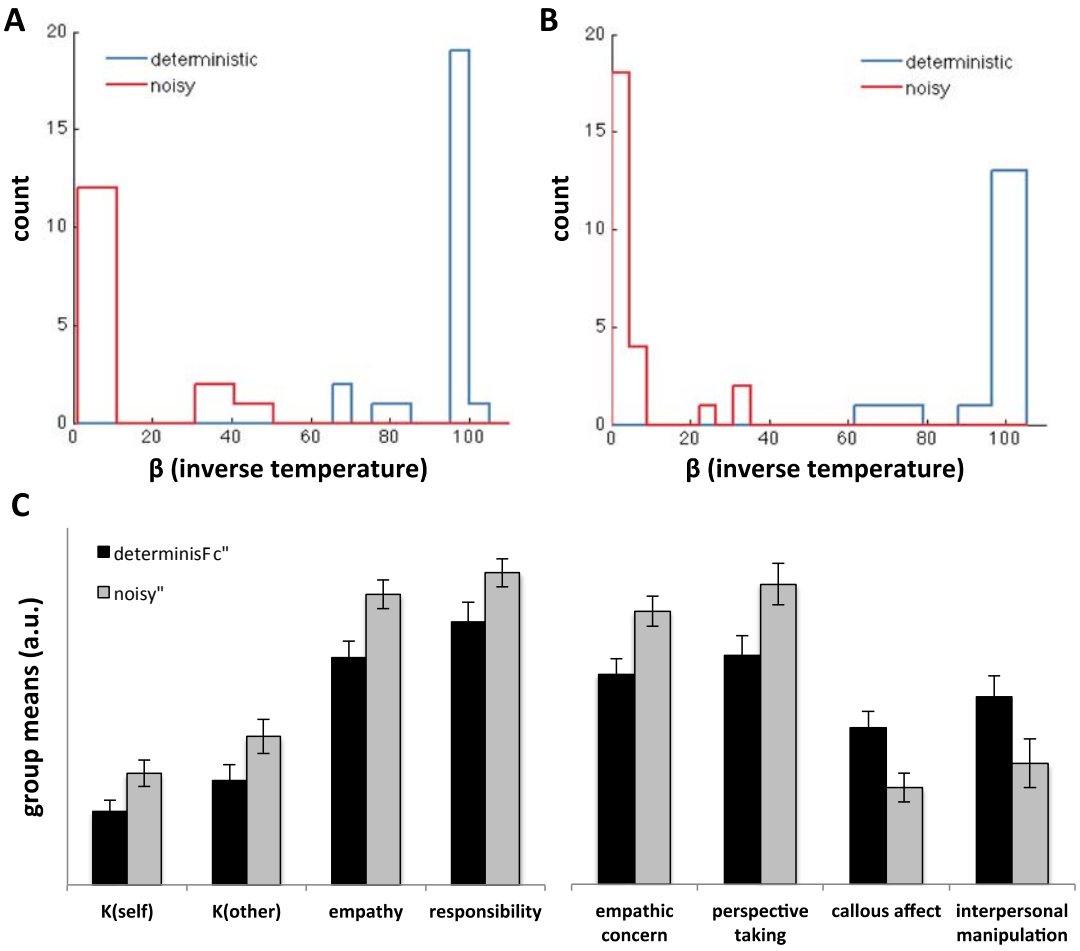
even if there is only a tiny difference in subjective value between the best and worst options; when it is low, choices will seem random. Meanwhile when irreducible noise is high, subjects may occasionally make "irrational" choices where a very unattractive option is selected over a very attractive one.

Previous work has linked the choice accuracy parameter to subjective confidence in choice. Subjects were asked to make decisions between pairs of food items that differed in attractiveness. After each choice they were asked to rate how confident they felt about their choice. Low-confidence choices were significantly noisier than high-confidence choices, as indicated by the choice accuracy parameter (De Martino, Fleming, Garrett, & Dolan, 2013). This suggests that subjective feelings of confidence in choice are related to objectively quantifiable aspects of decisions that are captured in the choice model.

In our experiments we investigated whether hyperaltruism was related to the noisiness of choices for self and others. If people are hyperaltruistic because they are more uncertain when deciding for others relative to themselves, we should see that the degree of hyperaltruism correlates with the extent to which choices are noisier for others than self, as indicated by either the choice accuracy parameter or the irreducible noise parameter. We indeed observed this effect for the choice accuracy parameter (Fig 3A). Hyperaltruistic subjects had noisier choices for others than for self. Although we did not measure subjective confidence in our studies, the findings of De Martino et al. (2013) would suggest that hyperaltruism is related to less confidence when choosing for others than for self.

Another interesting observation is that two distinct groups of subjects could be segregated with respect to choice accuracy in both studies (Fig. 3B-C). When we pooled this data and examined group differences it became evident that the group with noisier choices displayed more prosocial characteristics than those whose choices were deterministic (Fig. 3D). Noisier subjects were more harm averse than the deterministic subjects, both for themselves. Noisier subjects also reported feeling more empathy and responsibility for the Receiver than did the deterministic subjects. Finally, noisier subjects possessed more empathic traits (perspective taking and empathic concern) and fewer psychopathic traits (callous affect and interpersonal manipulation) than the deterministic subjects.

**Figure 3. Choice noisiness is associated with prosocial traits and behavior.**
(A,B) In both studies we observed a bimodal distribution of the choice accuracy parameter. We split subjects into groups with high parameter estimates ('deterministic') and low parameter estimates ('noisy'). (C) The noisy and deterministic groups differed on task performance (left panel) and personality traits (right panel); a.u.=arbitrary units.




Together these data provide further evidence that moral preferences are associated with uncertainty when decisions impact others. Intriguingly, however, this perspective is inconsistent with a separate line of research demonstrating increased selfish behavior in the face of uncertainty. In one study, Dana and colleagues compared choices in a standard dictator game with those in a modified dictator game where the outcome for the recipient was uncertain (Dana, Weber, & Kuang, 2007). Across all conditions dictators chose between two options ("A" and "B"). In a baseline treatment, 74% of dictators preferred option B ($5-$5) to option A ($6-$1). In a "hidden information" treatment, dictators could again receive $6 for choosing A and $5 for choosing B. However, they did not

know initially whether choosing A or B would yield $1 or $5 for the receiver. In fact the payoffs could be (A: $6,$1; B: $5,$5), as in the baseline treatment, or instead (A: $6,$5; B: $5,$1). The actual outcomes were determined by a coin flip and could be costlessly revealed before the dictators made their decision. Here, only 47% of dictators revealed the true payoffs and selected the fair option – a significantly lower proportion than in the baseline treatment. This suggests that fair choices in the baseline treatment are at least partially motivated by a desire to appear fair rather than a true preference for fair outcomes. A second study showed similar results when the outcome for the recipient was determined jointly by two dictators. In this study, a selfish choice by one dictator could not guarantee a bad outcome for the recipient, as the other dictator could still ensure a fair outcome. Thus, when uncertainty about outcomes obscures the relationship between choices and consequences, more people choose selfishly. In other words, uncertainty provides a smokescreen behind which selfishness can hide.

Another study suggests that not only does uncertainty promote selfishness, but people actually prefer more uncertainty rather than less in the context of social dilemmas. Haisley and Weber (2010) compared choices in a "risky" dictator game with those in a more uncertain "ambiguous" dictator game. In the risky game, the prosocial choice yielded $2 for self and $1.75 for the receiver, and the selfish choice yielded $3 for self and either 0 or $0.50 for the receiver (each occurring with 50% probability). In the ambiguous game, the prosocial choice again yielded $2 for self and $1.75 for the receiver, and the selfish choice yielded $3 for self and either 0 or $0.50 for the receiver, each occurring with an *unknown* probability. Thus the ambiguous game contained more uncertainty about the outcome of the receiver. Dictators were more likely to choose the

selfish option in the ambiguous game, and this was driven by self-serving beliefs about the likely outcome for the receiver. In other words, the increased uncertainty about outcomes in the ambiguous game allowed room for dictators to convince themselves that the selfish choice would not be too harmful for the receiver, and these self-serving beliefs led them to prefer more uncertainty rather than less (Haisley & Weber, 2010).

How can these findings be reconciled with our recent observation that hyperaltruism relates positively to uncertainty in choice? One possibility concerns the nature of the outcome for the receiver. Thus far all of the studies demonstrating increased selfishness in the face of uncertainty have investigated choices about monetary outcomes. However, as outlined above it is not clear whether selfish choices in these paradigms cause suffering for the receiver. It may be the case that uncertainty only increases altruism for truly *moral* decisions that concern the suffering of another person. In these kinds of decisions, moral risk aversion may drive increased altruism under uncertainty, whereas in monetary exchange decisions the desire to appear fair may primarily drive altruistic choices, and when appearances can be preserved under uncertain conditions, selfishness may prevail.

Alternatively, it may be that different kinds of uncertainty have different effects on altruism. People may be uncertain about *what the outcome will be,* or about *how the outcome will be experienced.* The studies with monetary outcomes involved the first kind of uncertainty, whereas our recent studies with painful outcomes involved the second kind. When outcomes are uncertain, the link between actions and outcomes is obscured and this may degrade the sense of moral responsibility, enabling selfish behavior. However, when outcomes are certain and responsibility is thus preserved, uncertainty

about how those outcomes will be experienced may lead to moral risk aversion. Teasing apart how different types of uncertainty affect moral decisions is an intriguing avenue for future research. One important question is how these different types of uncertainty are reflected in the parameters of the choice model. It may be that the choice accuracy parameter is differentially sensitive to uncertainty about outcomes, which can in principle be resolved with more information, versus uncertainty about experiences, which can never be resolved since the subjective experience of others is fundamentally unknowable (Harsanyi, 1955; Nagel, 1974).

**Moral decision-making as a learning process**

In the previous section I discussed how uncertainty about outcomes and the experiences of others affects moral decision-making. What if people are also uncertain about their own moral preferences? Thus far we have treated preferences as fixed quantities that are fully known to the decision-maker. From this perspective, decision-making simply involves translating the underlying subjective values into active choices. But if people are uncertain about their own preferences, the process of decision-making could be a form of learning whereby people discover their preferences by making choices and then observing their reactions to those choices. From this perspective, preferences are not fixed quantities but rather take the form of probabilistic belief distributions. This idea emerges from an "active inference" framework for decision-making that characterizes decision-making as a (Bayesian) inference problem (Friston et al., 2013, 2014). Inferred representations of self and others may serve the function of predicting and optimizing the potential outcomes of social interactions (Moutoussis, Fearon, El-Deredy, Dolan, &

Friston, 2014). This idea is also related to economic models of self-signaling, in which actions provide signals to ourselves that indicate what kind of people we are (Bodner & Prelec, 2003).

This perspective may shed light on various well-documented yet conflicting aspects of moral decision-making: moral licensing, conscience accounting, and moral consistency. Moral licensing describes the process whereby morally good behavior "licenses" subsequent immoral behavior (Merritt, Effron, & Monin, 2010). For example, laboratory studies have shown that subjects who purchase environmentally friendly products are subsequently more likely to cheat for personal gain (Mazar & Zhong, 2010), and subjects who demonstrate nonracist attitudes are more likely to subsequently show racist behavior (Monin & Miller, 2001). Moral licensing was also evident in a recent study of real-world moral behavior using ecological momentary assessment (Hofmann, Wisneski, Brandt, & Skitka, 2014). Conscience accounting describes the reverse process, whereby people who initially behave immorally are more likely to compensate for their bad behavior by doing a good deed. For instance, subjects who initially told a lie were more likely to donate to charity than those who did not lie (Gneezy, Imas, & Madarasz, 2012). Finally, moral consistency describes a tendency to make moral choices that are similar to previous ones. It is well known from work on cognitive dissonance that behaving inconsistency is uncomfortable (Festinger, 1962; Higgins, 1987); the "foot-in-the-door" persuasion technique capitalizes on this, making people more likely to help in the future if they have helped in the past (Freedman & Fraser, 1966). Related work on moral identity has shown that inducing people to recall past moral behavior increases the likelihood of future moral deeds (Shao, Aquino, & Freeman, 2008). Moral licensing and

conscience accounting seem to contradict moral consistency, as the former involve inconsistent patterns of choice. What could explain this contradiction?

If moral preferences are not fixed known quantities but rather belief distributions whose precision can be improved with experience, we might expect moral choices to initially be rather noisy in the absence of experience. This noise could manifest in the form of moral licensing and conscience accounting. With increased experience in the choice context, however, people should learn about their own preferences and choices should become less noisy, resulting in moral consistency. Treating moral decision-making as an inference problem leads to a prediction that moral licensing and conscience accounting should manifest in new contexts where decision-makers have limited experience, while moral consistency should appear in situations where decision-makers have extensive experience. Furthermore, computational models of choice should reflect a quantitative relationship between the precision of beliefs about one's own preferences and the noisiness of choices – with choices becoming less noisy over time as beliefs become more precise.

Considering moral decision-making as a learning process also highlights a possible role for prediction errors in guiding moral decisions. In standard reinforcement learning models, prediction errors represent discrepancies between expected and experienced outcomes, and guide learning by adjusting expectations. In the context of moral decisions, choices that are inconsistent with one's self-concept may similarly generate prediction errors. For example, if someone is uncertain about how much he dislikes cheating, and he predicts that he won't mind it much, but then after cheating he feels very guilty, the resulting prediction error teaches him something about his own

preferences -- he now knows he dislikes cheating with greater certainty than before. There is indeed evidence for prediction error-like signals during social decision-making (Chang, Smith, Dufwenberg, & Sanfey, 2011; Kishida & Montague, 2012; Xiang, Lohrenz, & Montague, 2013). Whether similar signals are present during moral decisions, and play a role in dynamic aspects of moral choices like licensing and consistency, is unknown.

The concept of prediction errors may also provide a computational account of the phenomenon of moral hypocrisy, where people view themselves as moral while failing to act morally (Batson, Kobrynowicz, Dinnerstein, Kampf, & Wilson, 1997). Batson et al. (1999) showed that moral hypocrisy can be reduced by heightening self-awareness, suggesting that moral hypocrisy arises when people fail to compare their behavior with their own moral standards (Batson, Thompson, Seuferling, Whitney, & Strongman, 1999). On a computational level, this might correspond to a suppression of prediction errors resulting from discrepancies between personal moral values and immoral behavior -- a hypothesis that could be tested with neuroimaging.

**Concluding remarks**

Research on value-based decision-making has shown that it is informative to examine not just the choices people make, but also the computational mechanisms that underlie *how* those decisions are made (De Martino et al., 2013; Krajbich, Armel, & Rangel, 2010). Recent work has extended this approach to investigating social and also moral decision-making (Crockett et al., 2014; Kishida & Montague, 2012). A model-based approach can provide additional insight into human morality by describing how cognitive processes

such as uncertainty and learning interact with and influence preferences. Another advantage of this approach is that it can generate novel and testable predictions about the dynamics of moral decision-making, such has how they unfold over time and how past decisions can influence future ones. Finally, computational models advance theory by forcing researchers to formalize the components of cognition and how they operate at an algorithmic level. This approach thus holds promise for addressing long-standing unanswered questions about human moral cognition and behavior.

**References**


Ashton, N. L., & Severy, L. J. (1976). Arousal and Costs in Bystander Intervention. *Personality and Social Psychology Bulletin*, *2*(3), 268–272. doi:10.1177/014616727600200313

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), 290–302. doi:10.1037/0022-3514.40.2.290

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), 1335–1348. doi:10.1037/0022-3514.72.6.1335

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), 525–537. doi:10.1037/0022-3514.77.3.525

Bernhardt, B. C., & Singer, T. (2012). The Neural Basis of Empathy. *Annual Review of Neuroscience*, *35*(1), 1–23. doi:10.1146/annurev-neuro-062111-150536

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

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. doi:10.1016/j.neuron.2011.02.056

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. doi:10.1162/003355302760193904

Clithero, J. A., & Rangel, A. (2013). The Computation of Stimulus Values in Simple Choice. In *Neuroeconomics: Decision Making and the Brain* (2nd ed., pp. 125–147). Academic Press.

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. doi:10.1073/pnas.1408988111

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

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. doi:10.1111/j.1467-9280.2006.01834.x

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. doi:10.1007/s00199-006-0153-z

Darley, J. M., & Latane, B. (1968). BYSTANDER INTERVENTION IN EMERGENCIES: DIFFUSION OF RESPONSIBILITY. *Journal of Personality and Social Psychology*, *8*(4, Pt.1), 377–383. doi:10.1037/h0025589

David, C., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem." *Emotion*, *12*(2), 364–370. doi:10.1037/a0025561

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision Making, Affect, and Learning: Attention and Performance XXIII*. Oxford University Press.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. doi:10.1038/nn.3279

Engel, C. (2011). Dictator games: a meta study. *Experimental Economics*, *14*(4), 583–610. doi:10.1007/s10683-011-9283-7

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A Bad Taste in the Mouth Gustatory Disgust Influences Moral Judgment. *Psychological Science*, *22*(3), 295–299. doi:10.1177/0956797611398497

Fehr, E., & Krajbich, I. (2013). Social Preferences and the Brain. In *Neuroeconomics: Decision Making and the Brain* (2nd ed.). Academic Press.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., & Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Social Cognitive and Affective Neuroscience*, *7*(7), 743–751. doi:10.1093/scan/nss069

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. doi:10.1016/j.cognition.2012.02.001

Festinger, L. (1962). *A Theory of Cognitive Dissonance*. Stanford University Press.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, *4*(2), 195–202. doi:10.1037/h0023552

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, *7*. doi:10.3389/fnhum.2013.00598

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: dopamine and decision-making. *Philosophical*

*Transactions of the Royal Society of London B: Biological Sciences*, *369*(1655), 20130481. doi:10.1098/rstb.2013.0481

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision Making and the Brain*. Academic Press.

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

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. doi:10.1037/a0021847

Gray, K., Waytz, A., & Young, L. (2012). The Moral Dyad: A Fundamental Template Unifying Moral Judgment. *Psychological Inquiry*, *23*(2), 206–215. doi:10.1080/1047840X.2012.686247

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124. doi:10.1080/1047840X.2012.651387

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. doi:10.1016/j.cognition.2009.02.001

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. doi:10.1126/science.1062872

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. doi:10.1037/0033-295X.108.4.814

Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science*, *316*(5827), 998–1002. doi:10.1126/science.1137651

Haisley, E. C., & Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, *68*(2), 614–625. doi:10.1016/j.geb.2009.08.002

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations. *Science*, *316*(5831), 1622–1625. doi:10.1126/science.1140738

Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010). Value Computations in Ventral Medial Prefrontal Cortex during Charitable Decision Making Incorporate Input from Regions Involved in Social Cognition. *The Journal of Neuroscience*, *30*(2), 583–590. doi:10.1523/JNEUROSCI.4089-09.2010

Harsanyi, J. C. (1955). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy*, *63*. Retrieved from http://econpapers.repec.org/article/ucpjpolec/v_3A63_3Ay_3A1955_3Ap_3A309.htm

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural Responses to Ingroup and Outgroup Members' Suffering Predict Individual

Differences in Costly Helping. *Neuron*, *68*(1), 149–160. doi:10.1016/j.neuron.2010.09.003

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, *94*(3), 319–340. doi:10.1037/0033-295X.94.3.319

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

Horberg, E. J., Oveis, C., & Keltner, D. (2011). Emotions as Moral Amplifiers: An Appraisal Tendency Approach to the Influences of Distinct Emotions upon Moral Judgment. *Emotion Review*, *3*(3), 237–244. doi:10.1177/1754073911402384

Kishida, K. T., & Montague, P. R. (2012). Imaging Models of Valuation During Social Interaction in Humans. *Biological Psychiatry*, *72*(2), 93–100. doi:10.1016/j.biopsych.2012.02.037

Korman, J., Voiklis, J., & Malle, B. F. (2015). The social life of cognition. *Cognition*, *135*, 30–35. doi:10.1016/j.cognition.2014.11.005

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. doi:10.1038/nn.2635

Leonhardt, J. M., Keller, L. R., & Pechmann, C. (2011). Avoiding the risk of responsibility by seeking uncertainty: Responsibility aversion and preference for indirect agency when choosing for others. *Journal of Consumer Psychology*, *21*(4), 405–413. doi:10.1016/j.jcps.2011.01.001

Mazar, N., & Zhong, C.-B. (2010). Do Green Products Make Us Better People? *Psychological Science*. doi:10.1177/0956797610363538

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral Self-Licensing: When Being Good Frees Us to Be Bad. *Social and Personality Psychology Compass*, *4*(5), 344–357. doi:10.1111/j.1751-9004.2010.00263.x

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), 33–43. doi:10.1037/0022-3514.81.1.33

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. doi:10.1016/j.concog.2014.01.009

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435. doi:10.2307/2183914

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial Behavior: Multilevel Perspectives. *Annual Review of Psychology*, *56*(1), 365–392. doi:10.1146/annurev.psych.56.091103.070141

Piliavin, I. M., Piliavin, J. A., & Rodin, J. (1975). Costs, diffusion, and the stigmatized victim. *Journal of Personality and Social Psychology*, *32*(3), 429–438. doi:10.1037/h0077092

Piliavin, J. A., & Piliavin, I. M. (1972). Effect of blood on reactions to a victim. *Journal of Personality and Social Psychology*, *23*(3), 353–361. doi:10.1037/h0033166

Shao, R., Aquino, K., & Freeman, D. (2008). Beyond Moral Reasoning: A Review of Moral Identity Research and Its Implications for Business Ethics. *Business Ethics Quarterly*, *18*(4), 513–540.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., … Sanchez-Vives, M. V. (2006). A Virtual Reprise of the Stanley Milgram Obedience Experiments. *PLoS ONE*, *1*(1), e39. doi:10.1371/journal.pone.0000039

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), 76–105. doi:10.1016/0022-1031(91)90011-T

Story, G. W., Vlaev, I., Seymour, B., Winston, J. S., Darzi, A., & Dolan, R. J. (2013). Dread and the Disvalue of Future Pain. *PLoS Comput Biol*, *9*(11), e1003335. doi:10.1371/journal.pcbi.1003335

Stürmer, S., Snyder, M., Kropp, A., & Siem, B. (2006). Empathy-Motivated Helping: The Moderating Role of Group Membership. *Personality and Social Psychology Bulletin*, *32*(7), 943–956. doi:10.1177/0146167206287363

Thomson, J. J. (1976). KILLING, LETTING DIE, AND THE TROLLEY PROBLEM. *The Monist*, *59*(2), 204–217.

Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, *12*(3), 579–590. doi:10.1037/a0024611

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology*, *44*(5), 1334–1338. doi:10.1016/j.jesp.2008.03.010

Vlaev, I. (2012). How different are real and hypothetical decisions? Overestimation, contrast and assimilation in social interaction. *Journal of Economic Psychology*, *33*(5), 963–972. doi:10.1016/j.joep.2012.05.005

Vlaev, I., Seymour, B., Dolan, R. J., & Chater, N. (2009). The price of pain and the value of suffering. *Psychological Science*, *20*(3), 309–317.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational Substrates of Norms and Their Violations during Social Exchange. *The Journal of Neuroscience*, *33*(3), 1099–1108. doi:10.1523/JNEUROSCI.1642-12.2013

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. doi:10.1073/pnas.0914826107

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(20), 8235–8240. doi:10.1073/pnas.0701408104

Zaki, J., & Mitchell, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences*, *108*(49), 19761–19766.