Moral values and motivations: How special are they?

Ryan Miller and Fiery Cushman

Are the hedonic and motivational properties of moral values handled by specialized systems in the brain, or might they be encoded in domain-general regions that process nonmoral rewards and punishments? To a large degree, we find that various aspects of moral value—including the subjective value of moral actions, outcomes, and their integration—are supported by a domain-general cognitive and neural architecture implicated in reward-related processes and economic decision-making.

Introduction

Moral values are central to human identity. Charles Darwin considered the human moral sense, or "conscience," to be the single most important attribute distinguishing us from other animals (1872), and recent research suggests that the particular constellation of moral traits you possess is a big part of what makes you "you" (Strohminger & Nichols, 2014). Given this privileged status, we ask a simple question: Is there something special that makes moral value different from other kinds of values that humans hold? Is the way it is acquired, stored, or implemented in the brain fundamentally special? Or might the difference be less sharp, with a common system (or systems) handling value of all kinds?

In many ways, the moral value we attach to particular behaviors or social outcomes, like generosity, honesty, or fairness, can be contrasted with other types of value, like a love of money, chocolate cake, or Mozart. We expect others to have particular moral values—and punish them when they don't. In contrast, we don't punish them for hating Mozart (at least not often). Similarly, we feel guilt or shame when our own actions are inconsistent with our moral values, but we don't usually feel guilt or shame when we violate our food preferences and try a new dish. Moral values also tend to be sacred, meaning people are unwilling to place a material price tag on them or openly trade them against secular goods (Tetlock, Kristel, Beth, Green, & Lerner, 2000).

A closer look, however, reveals broad similarities between moral and nonmoral value. Both motivate us to obtain certain goals or desirable outcomes—like the welfare of sick children or the newest technological gadget—and we experience pleasure in both cases when we succeed. Both types of value are also heavily

influenced by the specific culture in which we live; just as local customs shape our tastes in music, food, or beauty, they also shape how we view harm, fairness and charitable obligations (Henrich et al., 2001; Lamba & Mace, 2013). Our valuation of morally laden acts, like sacrificing one individual to save others, also appears to be susceptible to many of the same biases that plague the valuation of monetary goods during economic decision-making (Rai & Holyoak, 2010). And, as it happens, moral values *can* be traded off against each other and, with the right rhetorical gloss, even against material interests (Tetlock, 2003). Many individuals are also perfectly willing to bargain sacred values for monetary gain in practice (especially when they think no one is watching), as scandal-prone politicians often remind us.

Drawing on evidence from cognitive neuroscience, neuroeconomics, and social psychology, we argue that these similarities are more than superficial coincidences. Rather, they reflect a shared cognitive and neural architecture underlying moral and nonmoral value. This isn't to say, of course, that there is *nothing* special about moral values. The basic claim is that the *motivational*, and perhaps affective, aspects of moral value—those intrinsic feelings that make you *want* to help a charity and feel pleasure when you do, or avoid harming someone and feel bad when you don't—are encoded by a domain-general system that also represents and processes a host of nonmoral rewards and punishments. Furthermore, the process of moral learning, whereby we update the moral value that we assign to particular actions or behaviors, is likely to be supported by domain-general learning processes that have been consistently identified as important in learning the value of nonmoral goods and actions.

In the first part of this chapter, we'll look at four basic lines of research supporting this claim. First, we'll examine evidence that the subjective value (and disvalue) of morally relevant, prosocial (and antisocial) *outcomes* is encoded in the same brain regions as nonmoral rewards and punishments. We'll then consider how moral *action* values—like those placed on generosity or non-violence—might rely on the same cognitive and neural processes that support action valuation in nonmoral domains. Third, we'll look at how social reinforcers that are important to learning moral norms (like average group behavior or expressions of approval and disgust) appear to update value representations in the brain via the same processes as nonsocial rewards and punishments. Fourth, we'll consider how these values influence moral judgment, including research that they are traded against each other in a way that resembles economic decision-making. In the final section of the chapter, we'll highlight several important questions that should be addressed by future research.

Moral Value and Nonmoral Machinery

Outcome Value

The subjective values of a wide variety of pleasurable and aversive outcomes appear to be encoded in a common network of neural structures. The receipt of positive stimuli, like food, sex, and money, is most prominently associated with activity in the ventral striatum (VS) and the medial orbitofrontal cortex (mOFC; Bartra, McGuire, & Kable, 2013; Kable & Glimcher, 2007; Liu, Hairston, Schrier, & Fan, 2011). Activity in the VS has been found to correlate with self-reported ratings of pleasure (Salimpoor, Benovoy, Larcher, Dagher, & Zatorre, 2011; Sescousse, Li, & Dreher, 2015), and

activity in both the VS and mOFC during the passive viewing of items predicts subsequent choice of those same items (Levy, Lazzaro, Rutledge, & Glimcher, 2011). Aversive outcomes, on the other hand, are more often associated with activity in the anterior insula (AI) and anterior cingulate cortex (ACC). A large meta-analysis of reward-related studies found that the value of "negative rewards" (i.e. punishments) is preferentially encoded in the AI and ACC (Bartra et al., 2013), and activation in the AI correlates with the self-reported intensity of affective states (Zaki, Davis, & Ochsner, 2012). Both these regions are also central components of what has been dubbed the "pain matrix", a network of regions consistently involved in the subjective experience of pain (Davis, 2000).

Notably, a variety of social concerns (often referred to as 'social preferences'; Fehr & Fischbacher, 2002) are also encoded in these very same regions (see Ruff & Fehr, 2014 for a review). For instance, the way that we value others' well-being looks very similar to our own. When good things happen to others—especially if we like them, or if they're similar to us—we show increased activity in overlapping regions of the VS (Mobbs et al., 2009). Watching others in pain, on the other hand, is associated with activity in the AI and ACC (Jackson, Brunet, Meltzoff, & Decety, 2006; Lamm, Nusbaum, Meltzoff, & Decety, 2007; Singer et al., 2004), and the magnitude of AI response predicts the willingness to reduce an in-group member's pain by enduring pain oneself (Hein, Silani, Preuschoff, Batson, & Singer, 2010). Interestingly, and perhaps troublingly, the decision *not* to help an out-group member is best predicted by ventral striatal activity, suggesting that taking pleasure in others' pain may be an important inhibitor of prosocial action.

The moral value of fairness is associated with similar neural signatures. One of the tools most commonly used to study fairness preferences in the lab is the Ultimatum Game. In this game, one player, the Decider, is endowed with an initial sum of money, and she has to decide how much of it to share with a second player, the Responder. If the Responder doesn't like the offer, he can reject it, and neither player gets anything. Responders who are offered a fair share are more likely to accept the offer, feel happier about it, and show increased activity in both mOFC and VS (Tabibnia, Satpute, & Lieberman, 2008). Unfair offers lead to increased AI activation, and the magnitude of this neural response predicts rejection of the offer (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Of course, a fair offer is better than a low offer for the Responder, so preferences for fairness are necessarily confounded with self-interest in the Ultimatum Game.

A variety of other tasks provide even stronger evidence for the involvement of domain general valuation mechanisms in the pure fairness motive. Individuals who passively view a series of variable monetary allocations to themselves and another study participant show increased reward-related activity in mOFC and VS when the two totals are brought closer together, regardless of who is getting the money that turn (Tricomi, Rangel, Camerer, & O'Doherty, 2010). When making unilateral decisions about how to distribute money between themselves and others, participants show increased mOFC activity for equitable distributions, and increased AI activation when making inequitable distributions (Zaki & Mitchell, 2011). Furthermore, the AI response during these trials predicts overall unwillingness to make inequitable decisions. Finally, and perhaps most pertinent to a discussion on morality, disinterested third parties

making decisions about how to distribute money among *other* individuals also show insula activity when the proposed distribution is unfair, and this is related to both its rejection rate (Civai, Crescentini, Rustichini, & Rumiati, 2012; Corradi-Dell'Acqua, Civai, Rumiati, & Fink, 2013) and a willingness to pay money to create equality among the group (Dawes et al., 2012).

Researchers have also studied how the subjective value of mutual cooperation might be encoded in the brain. The standard method used in this literature is the Prisoner's Dilemma game, in which two partners are each privately faced with the decision to either 'cooperate' or 'defect'. Cooperation, which involves giving up a little so that your partner gains a lot, is obviously costly for the cooperator, but it leads to the greatest group benefit if both partners do it. On any given trial, however, an individual can do even better for herself if she defects (i.e. contributes nothing) while her partner cooperates. Out of the four possible combinations of cooperation and defection, mutual cooperation leads to the highest activity in reward-related regions (including the VS and mOFC; Rilling et al., 2002). Interestingly, finding out that the other person cooperated when you defected is associated with the lowest activity in these regions despite the fact that it provides the highest monetary payout, underscoring the power of social consequences to modulate neural representations of reward.

Action Value

In the previous section, we covered several instances in which the subjective value of morally relevant outcomes appears to be represented in domain-general regions that also process nonmoral rewards. Many times when we talk about moral values, however, we don't simply mean the value that we place on states of affairs out in the world, like whether two people have an equal amount of money, or whether a friend is experiencing pleasure or pain. Rather, we refer to the value (or disvalue) we place on particular *actions* with social consequences, like charitable giving, or not harming others. Is there reason to believe that these moral 'action values' might also supported by a more domain-general cognitive and neural architecture?

To answer this question, we should first clarify what nonmoral action values are, and how they might be derived by learning processes in a normal environment of rewards and punishments. An action value is a motivational construct that represents the expected future reward conditioned upon choosing an action, and it is often based on the reward history of prior choices in relevantly similar circumstances. For instance, if a rat receives cheese every time it presses a lever, it will come to assign a high action value to lever-pressing, and it will be more likely to press the lever in the future. (The magnitude of this value will, of course, depend on just how much the rat likes the cheese).

In environments where action—outcome contingencies are relatively stable, this type of learning is very useful and can lead to benefits in computational efficiency and speed during future decision-making. By storing values directly on actions, the actor doesn't need to reference an internal model of the relationships between actions and the particular outcomes they lead to. Instead, it simply performs the action with the highest value. For this reason, learning and decision-making programs that rely on cached action values are often referred to as 'model-free', whereas those that choose actions by searching over an internal model of the world are referred to as "model-based" (Dayan & Niv, 2008). There is good evidence that humans naturally employ both of types of decision-making (Daw, Gershman,

Seymour, Dayan, & Dolan, 2011), and they seem to be supported at least in part by dissociable neural systems (Gläscher, Daw, Dayan, & O'Doherty, 2010).

One interesting feature of action values is that, under the right circumstances, they can continue to influence behavior *even when the associated outcome is no longer valuable*. For instance, rats who learn to press a lever for food will continue to press the lever even after they're full, provided training has been extensive enough (Dickinson, 1985). Though this insensitivity to devaluation is typically discussed in the context of drug addiction and compulsive behaviors in humans (Gillan et al., 2014; Schwabe, Dickinson, & Wolf, 2011), it has also been demonstrated in healthy adults using a task analogous to rat devaluation paradigms (Tricomi, Balleine, & O'Doherty, 2009).

Intriguingly, we also see instances of a similar phenomenon in the domain of morality. Consider the act of charitable giving, which has been found to be driven by a mix of two motives (Andreoni, 1990). On the one hand, you may give money because you value the welfare of the charity, often referred to as altruistic giving. On the other hand, you may give money because you value (or derive utility from) the act of giving itself. This motive has been termed 'warm glow' because of the positive feelings it engenders in the giver. How do we know warm glow exists? Individuals feel better when they're actively giving the money themselves vs. passively transferring it (Harbaugh, Mayr, & Burghart, 2007), personal giving is not crowded out by external sources of aid (Eckel, Grossman, & Johnston, 2005), and individuals continue to give even when they know their donation is completely ineffectual (Crumpler & Grossman, 2008). This insensitivity to changes in the utility of the donation mirrors what we see in devaluation paradigms and hints at the influence of a positive 'action value' attached to charitable giving.

Where might this action value come from? One possibility is that being embedded in a generally cooperative society teaches us that prosocial, cooperative behavior is actually in our long-term best interest (Peysakhovich & Rand, 2015; Rand et al., 2014). By continually having our cooperative acts positively reinforced, we come to place a high value on prosocial action, just like the cheese-loving rat places a high value on lever-pressing. In a study testing this idea, participants who were first assigned to a cooperative environment in which it paid to be nice were subsequently more likely to donate money to an anonymous individual (with no possibility of reciprocation) than participants who were first assigned to a more competitive environment in which few people cooperated (Peysakhovich & Rand, 2015). This suggests that the high value these cooperative individuals learned to place on prosocial action in the first phase "spilled over" into the second phase, even though there was no longer any rational self-benefit.

We also find evidence of the same dissociation between action and outcome values in cases of aversion to antisocial action. Cushman and colleagues (2012) brought participants into the lab and asked them to perform several pseudo-violent actions, like slamming the head of a lifelike baby doll against a table, or hitting a realistic-looking artificial leg with a hammer. Despite knowledge that these actions could cause no harm, participants showed significant signs of physiological aversion (measured by peripheral vasoconstriction) when simply thinking about performing these actions. Furthermore, these physiological changes were greater than in either a control group who performed metabolically matched actions, or a witness group who watched

someone else perform the same actions. People also report that they would feel uncomfortable performing pseudo-violent actions in more natural contexts, such as stabbing a fellow actor in the neck with a retractable stage knife as part of a play (Miller, Hannikainen, & Cushman, 2014). These data suggest that the motoric properties of canonically violent actions (like hitting, stabbing, and shooting), which usually cause substantial harm, can acquire a negative value that is sufficient to trigger an aversive response even after the harmful outcome has been removed.

Although multiple theoretical accounts have emerged in recent years detailing how model-free learning algorithms might shape both prosocial (Gęsiarz & Crockett, 2015) and antisocial (Crockett, 2013; Cushman, 2013) behavior and moral judgment, few, if any, studies have attempted to directly compare the neural circuits involved in the model-free learning of both moral and nonmoral action values. Several studies do, however, suggest that the positive values attached to prosocial actions and the negatives value attached to antisocial actions are represented in many of the same reward-related brain regions that we've previously discussed. The warm glow associated with the prosocial act of giving to a charity, for instance, appears to be localized to the ventral striatum (Harbaugh et al., 2007). Studies on violent behaviors are a bit more difficult to interpret, in part because they have not isolated the violent act from its harmful outcomes. Nevertheless, one study found that the aversiveness of imagined harmful actions, like forcibly removing organs from a young child, was encoded in (mid-)insula and ACC, and functional connectivity analyses suggest that the information in ACC was passed to mOFC during moral judgment (Hutcherson, Montaser-Kouhsari, Woodward, & Rangel, 2015). In another study, the aversiveness of up-close-and-personal harmful actions tracked activity in the amygdala, and this appraisal was integrated into an overall moral value representation in the mOFC (Shenhav & Greene, 2014). The amygdala is important in learning to avoid negative outcomes (Delgado, Jou, LeDoux, & Phelps, 2009), and could here represent the learned association between violent actions and the harm that they typically cause (Blair, 2007). Future studies will be necessary to obtain a more fine-grained picture of how action and outcome values are independently represented in these regions.

Feedback and Learning

So far, we've talked about the various cognitive and neural substrates of moral value, but we haven't said much about the reinforcers that create or modify these values. Given the consistency of moral norms within cultures and variability of moral norms between cultures (Henrich et al., 2001; Lamba & Mace, 2013), one of the primary ways to learn the specific values of your culture is via social feedback. This is likely to come in one of two forms: prescriptive (involving direct signals of approval and disapproval) or descriptive (involving information about others' behavior). If moral values are encoded in domain-general regions, we might expect the feedback that comes from these two sources to operate over the same domain-general circuitry as nonmoral feedback.

Consistent with this hypothesis, faces signaling disapproval elicit activity in the ACC (Burklund, Eisenberger, & Lieberman, 2007), and this same region, along with the anterior insula, is activated in individuals who are subject to social exclusion (Eisenberger, Lieberman, & Williams, 2003). Facial expressions of disgust (another potential form of disapproval) also appear to amplify error processing in the ACC (Boksem,

Ruys, & Aarts, 2011). Indicators of social approval, on the other hand, are associated with increased activity in more reward-related regions, including the VS and mOFC (Jones et al., 2011).

Descriptive norms also have a powerful effect on behavior—thanks to the human desire to conform—and this influence can be seen playing out in the same brain regions. When one finds out their behavior or preferences match the group norm, it elicits activity in the VS; when they deviate from the norm, it leads to increased activity in the anterior insula and ACC (Wu, Luo, & Feng, 2016). Furthermore, the magnitude of response in these latter regions predicts the likelihood that the individual will change her behaviors or preferences to match those of the group (e.g. Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009; Zaki, Schirmer, & Mitchell, 2011). Interestingly, this change in preference is often accompanied by a commensurate change in reward-related striatal activity, suggesting conformity involves an updating of intrinsic preferences, rather than a superficial acquiescence to social pressures (Wu et al., 2016).

From Value to Judgment

We've discussed several ways in which the hedonic and motivational properties of morally relevant outcomes and actions mirror those of their nonmoral cousins, both cognitively and neurally. And, it is easy to see how these properties might promote moral behavior. Just as we are more likely to order an entrée that our brain finds pleasing, we are more likely to donate money to a charity if we find it intrinsically rewarding. But what about the relation of these prosocial values to moral *judgment*? Does a desire to act charitably towards others influence your judgment that it is morally required? Does an aversion to harm influence your judgment that it is morally prohibited?

To address this question, Shenhav and Greene (2010) asked participants to judge the moral acceptability of killing one individual in order to save others, varying both the number of people saved and the probability that they would die if nothing was done. Not only was the expected value of action (number saved X probability) encoded in the VS, but sensitivities to this value in the brain showed up as sensitivities in moral judgment. In other words, the more this reward-related region tracked the value of lives saved when reading scenarios, the more the participant incorporated the value into her ratings of acceptability, suggesting that reward was indeed modulating perceptions of wrongness. There is also evidence for action values (as opposed to outcome values) influencing moral judgment. In a previously mentioned study, Cushman and colleagues (2012) found that performing pseudo-violent (harmless) actions generated signs of aversive arousal. The magnitude of this physiological aversion also predicted how wrong participants thought it would be to kill one individual to save many others. Similarly, how uncomfortable you *think* it would make you to perform pseudo-violent actions predicts your condemnation of harmful actions, even when controlling for things like empathy and emotional reactivity (Miller et al., 2014).

Two recent neuroimaging studies have provided a window into how exactly these action and outcome values might be influencing moral judgment. In economic decision-making, the values of two or more goods have to be compared to each other in order to make a choice, but often their values are not on the same scale (e.g. choosing a cake now, or your health in 20 years). To perform this feat, the brain transforms these values into a 'common currency' that appears to

be encoded in mOFC (Chib, Rangel, Shimojo, & O'Doherty, 2009; Kable & Glimcher, 2007; Plassmann, O'Doherty, & Rangel, 2007). Interestingly, this same process seems to be occurring during moral judgment. Using tasks that pit an aversive action (like killing) against a utilitarian justification (like saving lives), Shenhav and Greene (2014) and Hutcherson and colleagues (2015) have found evidence that the appraisal values of each individual option, as well as the integrative moral judgment, are represented in mOFC in the moments before a judgment is made.

A common thread running through these studies is that they involve conflict, or competing moral concerns. We believe this may tell us something about the circumstances in which value (as a potentially affect-laden, motivational construct) is most likely to influence judgment. Many moral propositions, like "Murder is wrong" are likely to be stored in semantic memory and easily referenced. This is presumably why psychopaths are able to recognize simple moral violations, despite having reduced motivation to comply with them (Aharoni, Sinnott-Armstrong, & Kiehl, 2012; Blair, 1995). However, we may lack clear propositional knowledge concerning which moral rules are more important than others. In a situation where two moral norms—e.g. *do not kill* vs. *save lives the most lives*—are in competition, it might be necessary to reference the affective or motivational associations you have with each norm in order to render a judgment. It is precisely these circumstances where the judgments of psychopaths appear to diverge most from healthy adults (Koenigs, Kruepke, Zeier, & Newman, 2012).

Conclusion and Future Directions

The fingerprint of domain-general reward and valuation processes can be seen in several key components of moral cognition. The hedonic and motivational value attached to prosocial and antisocial outcomes, the actions that lead to them, and the social feedback that shapes them all seem to be reflected in regions that have been implicated in generic reward-learning tasks. Furthermore, the values of competing moral concerns appear to be translated into a 'common currency' in mOFC during moral judgment, just as we see in economic decision-making. This shared neural architecture may reflect the outsized role social cooperation plays in human fitness and survival. Cooperative ventures can lead to great personal benefits in both the short term and long term, and placing intrinsic value on prosocial actions may facilitate their success. Indeed, humans are extremely sensitive to whether their partners *want* to cooperate for its own sake, or whether they only do so after calculating the costs (Hoffman, Yoeli, & Nowak, 2015).

Several important questions remain, however, concerning the nature of these value representations. First, it is currently unclear the extent to which we can truly interpret activity in reward-related regions like ventral striatum as "intrinsic" valuation (or "private acceptance"), divorced from social expectations and pressures. Some studies looking at conformity-induced changes in these regions have favored this view (Berns, Capra, Moore, & Noussair, 2010; Klucharev et al., 2009; Zaki et al., 2011), but the evidence is mixed. Brain regions involved in theory of mind, for instance, can modulate value representations in the mOFC (Hare, Camerer, Knoepfle, O'Doherty, & Rangel, 2010; Strombach et al., 2015), and knowledge of others' presence can amplify activity in VS during charitable

donations (Izuma, Saito, & Sadato, 2008). These studies highlight the context-dependent nature of value construction, and demonstrate that anticipated social rewards (like reputation) might simultaneously contribute to reward-related activity in these regions. Future neuroimaging studies might consider using alternative techniques like multi-voxel pattern analysis (MVPA) to dissociate multiple sources of reward.

We also lack clear evidence on the degree of specialization for moral stimuli within the reward system. Though the bulk of research comparing social and nonsocial reward have found extensive overlap, a growing number of studies hint at some degree of regional specificity (Ruff & Fehr, 2014). For instance, the values of money (nonsocial) and erotic (social) stimuli are encoded in distinct regions of the mOFC (Sescousse, Redouté, & Dreher, 2010), and learning about the reliability of nonsocial cues vs. human advisors in predicting reward seems to rely on computationally similar yet neurally adjacent processing streams (Behrens, Hunt, Woolrich, & Rushworth, 2008). Few, if any, studies, however, have directly compared moral learning to nonsocial reward-learning, and even fewer have compared moral learning to nonmoral *social* learning. These are two areas that are ripe for investigation.

Finally, further research is needed to understand how exactly moral *action* values are learned. Some scholars have rightly questioned whether we have the requisite reinforcement history to form robust action values by personal experiential learning, particularly when it comes to relatively rare antisocial actions like hitting, stabbing, or shooting (Ayars, 2016; also similar to the "poverty of the stimulus" argument, see Mikhail, 2007).

There are several potential solutions to this problem. First, we can dynamically adjust learning rates, or how fast action values are updated, depending on perceived certainty of the outcome (Behrens, Woolrich, Walton, & Rushworth, 2007). In other words, actions that are known to reliably cause harm may acquire strong negative values after very little experience. Second, watching others, also known as observational or vicarious learning, activates the same neural pathways as first-hand experience and can be an efficient way of learning actions one is unlikely to perform oneself (Burke, Tobler, Baddeley, & Schultz, 2010; Olsson, Nearing, & Phelps, 2007). Third, instructional learning can lead to top-down modulation of reinforcement learning pathways (Doll, Jacobs, Sanfey, & Frank, 2009; Li, Delgado, & Phelps, 2011), resulting in neural responses that mirror first-hand learning. Lastly, mental simulation can play an important role in shaping action values (Gershman, Markman, & Ross, 2014); by using our model-based system to simulate various actions and their likely rewards and punishments, we can "train up" the cached action values in our model-free system so that learning occurs much more quickly and efficiently. Which of these explanations best describes how moral action values form is an open question, but we hope this outline provides several fruitful avenues for future research.

References

Aharoni, E., Sinnott-Armstrong, W., & Kiehl, K. A. (2012). Can psychopathic offenders discern moral wrongs? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology*, *121*, 484–497.

Andreoni, J. (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal*, *100*, 464–477.

Ayars, A. (2016). Can model-free reinforcement learning explain deontological moral judgments? *Cognition*, *150*, 232–242.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis

of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427.

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

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221.

Berns, G. S., Capra, C. M., Moore, S., & Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *NeuroImage*, *49*, 2687–2696.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, *57*, 1–29.

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

Boksem, M. A. S., Ruys, K. I., & Aarts, H. (2011). Facing disapproval: Performance monitoring in a social context. *Social Neuroscience*, *6*, 360–368.

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 14431–14436.

Burklund, L. J., Eisenberger, N. I., & Lieberman, M. D. (2007). The face of rejection: Rejection sensitivity moderates dorsal anterior cingulate activity to disapproving facial expressions. *Social Neuroscience*, *2*, 238–253.

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *The Journal of Neuroscience*, *29*, 12315–12320.

Civai, C., Crescentini, C., Rustichini, A., & Rumiati, R. I. (2012). Equality versus self-interest in the brain: Differential roles of anterior insula and medial prefrontal cortex. *NeuroImage*, *62*, 102–112.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2013). Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. *Social Cognitive and Affective Neuroscience*, *8*, 424–431.

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

Crumpler, H., & Grossman, P. J. (2008). An experimental test of warm glow giving. *Journal of Public Economics*, *92*, 1011–1021.

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

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*, 2–7.

Darwin, C. (1872). *The Descent of Man, and Selection in Relation to Sex*. D. Appleton.

Davis, K. D. (2000). The neural circuitry of pain as explored with functional MRI. *Neurological Research*, *22*, 313–317.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204–1215.

Dawes, C. T., Loewen, P. J., Schreiber, D., Simmons, A. N., Flagan, T., McElreath, R., … Paulus, M. P. (2012). Neural basis of egalitarian behavior. *Proceedings of the National Academy of Sciences*, *109*, 6479–6483.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*, 185–196.

Delgado, M. R., Jou, R. L., LeDoux, J., & Phelps, L. (2009). Avoiding negative outcomes: tracking the mechanisms of avoidance learning in humans during fear conditioning. *Frontiers in Behavioral Neuroscience*, *3*. doi:10.3389/neuro.08.033.2009

Dickinson, A. (1985). Actions and Habits: The Development of Behavioural Autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *308*, 67–78.

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94.

Eckel, C. C., Grossman, P. J., & Johnston, R. M. (2005). An experimental test of the crowding out hypothesis. *Journal of Public Economics*, *89*, 1543–1560.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does Rejection Hurt? An fMRI Study of Social Exclusion. *Science*, *302*, 290–292.

Fehr, E., & Fischbacher, U. (2002). Why Social Preferences Matter – the Impact of Non-Selfish

Motives on Competition, Cooperation and Incentives. *The Economic Journal*, *112*, C1–C33.

Gershman, S. J., Markman, A. B., & Ross, A. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*, 182–194.

Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in Behavioral Neuroscience*, 135.

Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., … Robbins, T. W. (2014). Enhanced Avoidance Habits in Obsessive-Compulsive Disorder. *Biological Psychiatry*, *75*, 631–638.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*, 585–595.

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

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. *Journal of Neuroscience*, *30*, 583–590.

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*, 149–160.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review*, *91*, 73–78.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences*, *112*, 1727–1732.

Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and Utilitarian Appraisals of Moral Dilemmas Are Encoded in Separate Areas and Integrated in Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, *35*, 12593–12605.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of Social and Monetary Rewards in the Human Striatum. *Neuron*, *58*, 284–294.

Jackson, P. L., Brunet, E., Meltzoff, A. N., & Decety, J. (2006). Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain. *Neuropsychologia*, *44*, 752–761.

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., … Casey, B. J. (2011). Behavioral and Neural Properties of Social Reinforcement Learning. *Journal of Neuroscience*, *31*, 13039–13045.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*, 1625–1633.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement Learning Signal Predicts Social Conformity. *Neuron*, *61*, 140–151.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*, 708–714.

Lamba, S., & Mace, R. (2013). The evolution of fairness: explaining variation in bargaining behaviour. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*, 20122028.

Lamm, C., Nusbaum, H. C., Meltzoff, A. N., & Decety, J. (2007). What Are You Feeling? Using Functional Magnetic Resonance Imaging to Assess the Modulation of Sensory and Affective Responses during Empathy for Pain. *PLOS ONE*, *2*, e1292.

Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2011). Choice from Non-Choice: Predicting Consumer Preferences from Blood Oxygenation Level-Dependent Signals Obtained during Passive Viewing. *Journal of Neuroscience*, *31*, 118–125.

Li, J., Delgado, M. R., & Phelps, E. A. (2011). How instructed knowledge modulates the neural systems of reward learning. *Proceedings of the National Academy of Sciences*, *108*, 55–60.

Liu, X., Hairston, J., Schrier, M., & Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *35*, 1219–1236.

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

Miller, R., Hannikainen, I., & Cushman, F. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*. doi:10.1037/a0035361

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A. J., … Dalgleish, T. (2009). A Key Role for Similarity in Vicarious Reward. *Science*, *324*, 900–900.

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: the neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*, *2*, 3–11.

Peysakhovich, A., & Rand, D. G. (2015). Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory. *Management Science*, *62*, 631–647.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *Journal of Neuroscience*, *27*, 9984–9988.

Rai, T. S., & Holyoak, K. J. (2010). Moral Principles or Consumer Preferences? Alternative Framings of the Trolley Problem. *Cognitive Science*, *34*, 311–321.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*, 3677.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A Neural Basis for Social Cooperation. *Neuron*, *35*, 395–405.

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

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience*, *14*, 257–262.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science*, *300*, 1755–1758.

Schwabe, L., Dickinson, A., & Wolf, O. T. (2011). Stress, habits, and drug addiction: A psychoneuroendocrinological perspective. *Experimental and Clinical Psychopharmacology*, *19*, 53–63.

Sescousse, G., Li, Y., & Dreher, J.-C. (2015). A common currency for the computation of motivational values in the human striatum. *Social Cognitive and Affective Neuroscience*, *10*, 467–473.

Sescousse, G., Redouté, J., & Dreher, J.-C. (2010). The Architecture of Reward Value Coding in the Human Orbitofrontal Cortex. *Journal of Neuroscience*, *30*, 13095–13104.

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

Shenhav, A., & Greene, J. D. (2014). Integrative Moral Judgment: Dissociating the Roles of the Amygdala and Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, *34*, 4741–4749.

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*, 1157–1162.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*, 159–171.

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, *112*, 1619–1624.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The Sunny Side of Fairness: Preference for Fairness Activates Reward Circuitry (And Disregarding Unfairness Activates Self-Control Circuitry). *Psychological Science*, *19*, 339–347.

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

Tetlock, P. E., Kristel, O. V., Beth, S., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*, 853–870.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, *29*, 2225–2232.

Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, *463*, 1089–1091.

Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *71*, 101–111.

Zaki, J., Davis, J. I., & Ochsner, K. N. (2012). Overlapping activity in anterior insula during interoception and emotional experience. *NeuroImage*, *62*, 493–499.

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*, 19761–19766.

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological Science*, *22*, 894–900.