ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
2016, Vol. 25(2) 85–90
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0963721415624012
cdps.sagepub.com
SAGE

How Formal Models Can Illuminate Mechanisms of Moral Judgment and Decision Making

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

Abstract

The cognitive and affective processes that give rise to moral judgments and decisions have long been the focus of intense study. Here, I review recent work that has used mathematical models to formally describe how features of moral dilemmas are transformed into decisions. Formal models have traditionally been used to study perceptual and value-based learning and decision making, but until recently they had not been applied to the study of moral psychology. Using examples from recent studies, I show how formal models can provide novel and counterintuitive insights into human morality by revealing latent subcomponents of moral decisions, improving prediction of moral behavior, and bridging moral psychology and moral neuroscience.

Keywords
morality, decision making, judgment, altruism, formal models

Moral decisions often require trading off personal benefits against the welfare of others. How do we resolve conflicts between profit and harm? How do we judge others faced with similar dilemmas? And how are these processes implemented in the brain? These are some of the many questions moral psychologists and neuroscientists have explored over the past decade. Here, I suggest we can accelerate progress in moral psychology and neuroscience by applying formal algorithmic frameworks typically used to study perceptual and value-based learning and decision making (Love, 2015; Marr, 1982). This approach involves specifying mathematical models that describe in a precise, quantitative way how features of a choice problem are transformed into a decision. Recent studies have used this approach to describe moral algorithms—that is, how features of moral dilemmas (e.g., costs to the self, benefits to another) are transformed into moral judgments and decisions.

To illustrate this approach by analogy, imagine you want to bake a cake. The crucial first step is determining what ingredients are necessary for the cake—flour, sugar, eggs, milk, and so on. Next, you would need to know the amounts of each ingredient, and in what order to mix them. Similarly, past research in moral psychology has focused primarily on the critically important first step of identifying the key ingredients of moral judgments and decisions—norms, empathy, intentions, actions, outcomes, and so on (Batson, Duncan, Ackerman, Buckley, & Birch, 1981; Cushman, Young, & Hauser, 2006; Greene, 2014b; Malle, Guglielmo, & Monroe, 2014; Mikhail, 2007). Now that many of these ingredients have been identified, future work can begin to develop formal mathematical models that describe how they are combined, in what amounts and in what temporal order, to produce moral judgments and decisions.

Formal models can advance moral psychology in several ways. First, they can reveal latent subcomponents of moral decisions that would not otherwise be apparent from behavioral observation alone, and thereby advance psychological theories of morality. Second, by assigning numerical values (called *parameters*) to different subcomponents of decisions, formal models can improve prediction of behavior, which has clear value for applied settings. Finally, formal models bridge moral psychology and moral neuroscience by addressing specific mechanistic questions about neural activity and predicting how changes in the brain should affect behavior. In this way,

Corresponding Author:
Molly J. Crockett, Department of Experimental Psychology, University of Oxford, 9 South Parks Rd., Oxford OX1 3UD, United Kingdom
E-mail: molly.crockett@psy.ox.ac.uk

moral neuroscience can tackle the long-standing question of whether moral decisions are different from other kinds of decisions.

Formal Models Reveal Latent Subcomponents of Moral Decisions

How do people decide whether to harm others for personal gain? This question has long been the focus of experimental research (FeldmanHall et al., 2012; Masserman, Wechkin, & Terris, 1964), but a mechanistic understanding of moral decision making has been limited by the cognitive opaqueness of measured behaviors, such as choice proportions and response times. Formal models can identify latent subcomponents of moral decisions and describe how they interact to produce moral behavior. We recently adopted this approach in an experiment where participants made choices between different amounts of money and different numbers of painful electric shocks directed toward either themselves or an anonymous person (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014). We built mathematical models to relate objective features of the choice options (here, amounts of money and shocks) to their underlying subjective values (Fig. 1a). This allowed us to formally quantify several latent components of choice, including the negative utilities people ascribe to harming themselves and others (*harm aversion*). Strikingly, our model revealed that for most people, harming others is subjectively worse than harming oneself, and accordingly, most people were willing to sacrifice more money to avoid harming others than themselves (Crockett et al., 2014). Thus, by modeling latent components of choice, we were able to marshal empirical support for the proposal that moral behavior arises from a higher valuation of outcomes associated with better outcomes for others (Buckholtz, 2015).

Our model also exposed a novel cognitive process that relates to moral behavior. People vary in the extent to which they choose the more highly valued option—that is, their decision process is "noisy" and they sometimes make mistakes. This noise is another latent component of choice quantified within the model. We explored the possibility that people make noisier choices when deciding for others relative to themselves and that this would relate to moral behavior. Indeed, the extent to which people made noisier choices for others than for themselves was positively correlated with moral behavior (Crockett et al., 2014).

Formal models can also make precise predictions about the relationship between decisions and response times, which bears on contemporary debates about the automatic versus controlled nature of moral cognition (e.g., Crockett, 2013; Cushman, 2013). One popular class of models describes decision making as a process whereby a noisy signal indicating the relative value difference between two options dynamically evolves over time, and a choice is made when enough evidence has accumulated for one of the options (Ratcliff & McKoon, 2008). In recent studies, such models have afforded novel insights into how participants divide money between themselves and others (Hutcherson, Bushong, & Rangel, 2015) and decide to contribute money in public-goods games (Krajbich, Bartling, Hare, & Fehr, 2015). The models capture a number of latent variables, including the relative weights placed on payoffs for self and others and how much evidence favoring one choice option is required before a decision is made (Fig. 1b).

These models have revealed a number of surprising insights about altruistic behavior and the relationship between prosocial choice and response times. First, generous choices are slower if the weight placed on payoffs to oneself is higher, but faster if the weight placed on payoffs to others is higher. Thus, prosocial decisions are not always faster than selfish ones; the relationship between prosocial decisions and response times can even be manipulated by changing the costs of prosocial decisions (Krajbich et al., 2015). Second, when less evidence favoring one choice option is required before making a decision, generosity increases.[1] Thus, differences in generosity observed across individuals or contexts may not necessarily reflect differences in preferences, but could instead reflect differences in the noisiness of decisions. This has important implications for interpreting the effects of manipulations thought to influence decision noise, such as time pressure or cognitive load. The fact that amplified decision noise can increase generosity relates to a third insight: A substantial proportion of generous choices may be "mistakes" rather than reflecting true preferences (Hutcherson et al., 2015).

Formal Models Improve Prediction of Moral Behavior

By assigning numerical values to latent subcomponents of decisions, formal models can predict choices in new cases different from those used to estimate the original model. To illustrate this point in the domain of emotion, psychological research has observed that "bad is stronger than good" (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). Formal models of loss aversion quantify this effect and tell us that bad is about twice as subjectively strong as good (Kahneman & Tversky, 1979), which has an obvious predictive advantage over merely knowing that bad is stronger than good. Consider the set of gambles displayed in Table 1 and try to predict whether the average person would take each gamble, armed with either the descriptive theory or the formal model. For the first gamble (a 50/50 chance of gaining vs. losing $100), both

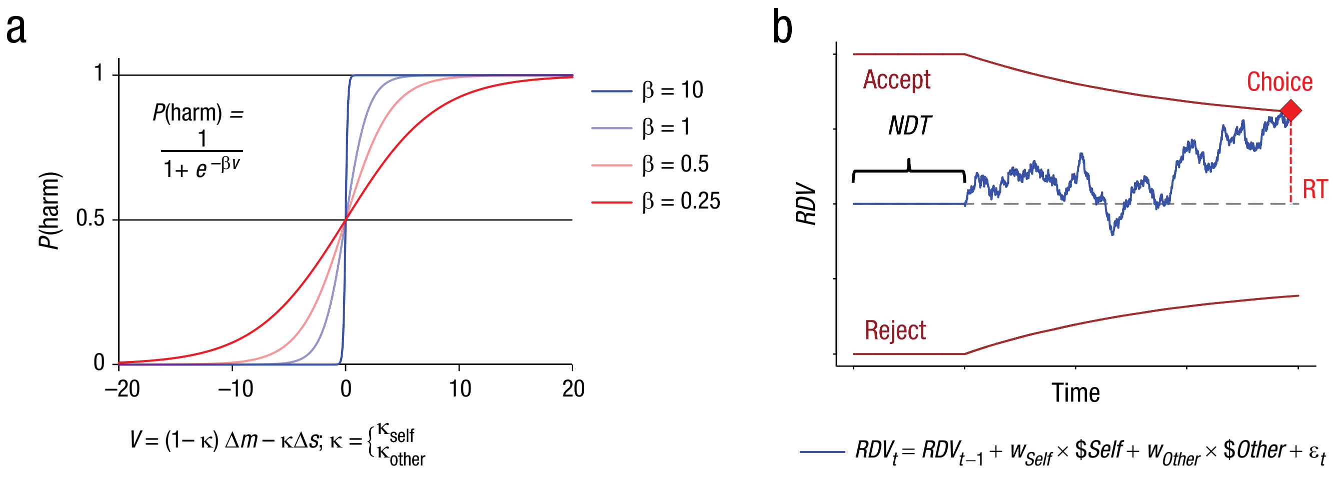

Fig. 1. Formal models of moral decision making. In a study by Crockett, Kurth-Nelson, Siegel, Dayan, and Dolan (2014), participants decided whether to gain money by delivering painful electric shocks to either themselves or others. Figure 1a depicts the probability of choosing the more harmful option (P(harm)) as a softmax function of its subjective value (V) relative to a default option. Choice functions are depicted for several values of the inverse temperature parameter, β, which determines the noisiness of choices. The subjective value of the more harmful option is modeled as a function of the amount of money gained (Δm) and the number of shocks delivered (Δs). The money and shock terms are scaled by a harm-aversion parameter (κ) that quantifies the exchange rate between money and pain and takes on different values for pain to self and others (κ_{self} and κ_{other}, respectively). In a study by Hutcherson, Bushong, and Rangel (2015), participants decided how to split a pot of money between themselves and others. As illustrated by Figure 1b, choices are made through the noisy accumulation of a relative decision value (*RDV*), based on a weighted sum of the amounts of money for self and other available on each trial. A response occurs when this accumulated value signal crosses a threshold, with a response time (RT) equal to the total accumulated time plus a nondecision time (*NDT*) to account for sensory and motor-related processes unrelated to the comparison process itself. Figure 1b was provided by Cendri Hutcherson.

the descriptive theory and the formal model predict the average person will not take the gamble. However, for the rest of the gambles, the formal model can predict behavior where the descriptive theory cannot. Moreover, the model makes quantitative predictions in terms of choice probabilities. Such predictions afford more stringent tests of theories because they are easier to falsify as measurement becomes more precise (Meehl, 1967).

Formal models allow us to test the intriguing possibility that unifying principles characterize decision making across a variety of domains, from perceptual decisions such as navigating through a parking lot, to simple value-based decisions such as selecting items from a restaurant menu, to moral decisions such as charitable giving. If this is the case, then formal models for decision making might generalize across contexts, enabling *out-of-sample predictions*—where the parameters extracted from one context (e.g., food choice) can predict decisions in a totally different context (e.g., sharing money). This was explicitly tested in a recent study. Krajbich and colleagues estimated the parameters of a decision model in a group of participants choosing among different foods. These parameters were then used to predict the choices of four different groups of participants making moral decisions about how to share money with an anonymous person or to punish others who treated them unfairly. Remarkably, the parameters estimated from food choices were able to predict moral decisions and response times with more than 90% accuracy (Krajbich, Hare, Bartling, Morishima, & Fehr, 2015). Thus, initial evidence suggests there may indeed be unifying principles of decision making that can be captured by formal models and used to make out-of-sample predictions. This has clear legal and policy implications: If researchers can discover a set of

Table 1. Toy Example Illustrating the Relative Predictive Power of Descriptive Theories Versus Formal Models

Gamble	Descriptive theory prediction	Formal model prediction
50% gain \$100, 50% lose \$100	Don't gamble	P(gamble) = 0%
50% gain \$100, 50% lose \$75	?	P(gamble) = 0%
50% gain \$100, 50% lose \$50	?	P(gamble) = 50%
50% gain \$100, 50% lose \$25	?	P(gamble) = 100%

Note: Participants are given the option to take a gamble with a 50% chance of gaining \$100 and a 50% chance of losing some amount. The formal model computes the probability of taking the gamble as a softmax transformation* of the expected value of the gamble, which is equal to the probability-weighted average of the possible outcomes, with the negative outcome multiplied by a factor of two (i.e., "bad is twice as strong as good").
*$P(\text{gamble}) = 1 / (1 + e^{-\beta V})$; β = decision noise parameter (here set to 1); V = expected value of gamble.

parameters that can describe decision making across a variety of contexts, these could be used to design incentive schemes that promote moral behavior and discourage antisocial behavior.

Practitioners in this area do not claim to have discovered universal parameters that can predict behavior in any situation. Rather, it is useful to think of this work as uncovering "baseline" parameters capturing the fact that, in aggregate, most people will exert a similar amount of effort to make decisions that satisfy their current goals (Krajbich et al., 2015). Undoubtedly, situational factors such as time pressure or distraction will modulate these parameters, and future research can evaluate whether they do so in systematic ways. Nevertheless, formal models provide a common mathematical language that can be used to compare decision processes across contexts and make out-of-sample predictions.

Formal Models Bridge the Moral Mind and Brain

Research on the neural basis of morality has identified a network of brain regions that are consistently activated during moral judgments and decisions (e.g., FeldmanHall et al., 2012; Greene, 2014a; Shenhav & Greene, 2010). Many of these same regions are involved in making decisions about outcomes for oneself, raising the critical question of whether there is anything "special" about moral decisions relative to other types of decisions (Cushman, 2015; Greene, 2014a). Another criticism relates to the issue of reverse inference, a logical fallacy where psychological processes are inappropriately inferred from activity in a particular brain region—for instance, when feelings of disgust are inferred from observing activity in the insula (Poldrack, 2011). Accordingly, there remains a great deal of skepticism as to whether neuroscience can tell us anything useful about morality, or indeed social cognition in general.

Formal models can advance moral neuroscience by precisely specifying the computations served by brain regions and neuromodulators during moral decision making. This mechanistic specificity sidesteps the issue of reverse inference and tightens the observed links between changes in the brain and changes in behavior. For example, the first study examining neuromodulation of moral judgments found that pharmacologically enhancing serotonin function reduced the judged permissibility of killing one person to save many others (Crockett, Clark, Hauser, & Robbins, 2010). The authors speculated that this effect reflected a role for serotonin in harm aversion: computing the negative utility of harming others. A recent study used formal models to test this hypothesis directly and found strong evidence that serotonin increases harm aversion (Crockett et al., 2015). Notably, the effect of the serotonin drug on harm aversion as quantified by the model was much stronger than the drug's previously reported effect on moral judgments. Because serotonin is thought to modulate value computations (Dayan & Huys, 2009), it is appropriate that the drug would more strongly influence the model's estimates of harm aversion—which reflect those computations directly—than moral judgments, which are thought to be the output of multiple valuation systems (Crockett, 2013; Cushman, 2013).

Formal models have also clarified the role of the insula in detecting and responding to violations of social norms, such as unfair treatment. Initial work in this area produced inconsistent findings, with some studies showing a positive correlation between insula responses and unfairness (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Tabibnia, Satpute, & Lieberman, 2008) and others showing a negative correlation (Dawes et al., 2012; Hsu, Anen, & Quartz, 2008; Wright, Symmonds, Fleming, & Dolan, 2011). Recent work inspired by computational models of learning has helped to resolve this inconsistency by suggesting that being treated unfairly could generate a prediction error—that is, a discrepancy between expectations and outcomes (Chang & Sanfey, 2013; Montague & Lohrenz, 2007). These models predict that neural responses to unfairness will be sensitive to people's expectations about what is fair, which could vary depending on the particulars of an individual experiment and thus potentially explain the discrepant findings across studies. Recent work has tested this explicitly by manipulating the average level of fairness experienced during the experiment and formally modeling the development of fairness expectations over time. In line with predictions, insula responses correlated with deviations from expected fairness norms (Chang & Sanfey, 2013; Xiang, Lohrenz, & Montague, 2013). This work can help us to better understand how people learn and use cultural norms to interact appropriately with others.

Because formal models provide a unifying framework that connects diverse kinds of decision making (Krajbich et al., 2015), they can be especially useful in addressing the question of whether moral decisions are different from other kinds of decisions. Initial work on this topic has highlighted commonalities between the neural networks subserving moral and amoral decisions (Hutcherson et al., 2015; Shenhav & Greene, 2010). However, moral values seem to constrain our behavior in ways that other kinds of values do not (Cushman, 2015). Resolving this issue will require studies that simultaneously investigate the neural basis of moral and amoral decisions within a common algorithmic framework. Such studies can more directly address whether moral decisions involve computations that are unique to moral cognition, or the same set of computations that are applied in decision making more generally (Ruff & Fehr, 2014), and reveal the extent to which moral judgments make use of normative (e.g., Bayesian) algorithms (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015).

Conclusion

The Nobel laureate and physicist Richard Feynman famously wrote, "What I cannot create, I do not understand" (cf. Agapakis, 2013). Formal models describing moral judgment and decision making represent a first step toward recreating, and therefore understanding, the cognitive processes that guide moral behavior. Imagine trying to program a robot to have humanlike moral values. Such a task would be impossible without formal mathematical models that describe, in numeric terms, how inputs to moral choices are transformed into outputs. Simply providing the robot with a set of if-then rules tailored to specific situations would be intractable because the robot might find itself in an infinite number of situations.

No single model can provide a definitive and unifying mechanism for moral decision making. Nor can the parameters derived from a single study serve as the final word on the numerical weights that apply to various components of moral decisions. Nevertheless, the advantage of formal models is that that they provide a common mathematical language that can be used to compare effect sizes across studies. As more and more studies apply these common frameworks, by aggregating their findings we can begin to formulate recipes that describe how to combine the ingredients of moral judgments and decisions. It may be the case that a relatively small number of models can capture most aspects of moral judgment and decision making. Alternatively, the richness and complexity of human morality may be impossible to boil down into a manageable set of mathematical equations. But we won't find out unless we try, and we will undoubtedly learn a lot in the process.

Recommended Reading

Brown, J. W. (2014). The tale of the neuroscientists and the computer: Why mechanistic theory matters. *Frontiers in Neuroscience*, *8*, Article 349. A highly accessible exposition of how formal models advance our understanding of neuroscience.

Cushman, F. (2015). (See References). A cogently argued analysis of similarities and differences between moral decisions and other kinds of value-based decisions.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). (See References). A recent study elegantly demonstrating how a model relating subjective value and response time can reveal surprising insights about the nature of generosity and its neural basis.

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). (See References). A remarkable recent study showing how parameters derived from food choices in one population can be used to predict, with high accuracy, moral choices in a different population.

Ruff, C. C., & Fehr, E. (2014). (See References). A comprehensive recent review of the neural basis of social and moral decision making.

Acknowledgments

I thank Peter Dayan, Cendri Hutcherson, Joshua Buckholtz, Ian Krajbich, Andreas Kappes, Tobias Gerstenberg, Walter Sinnott-Armstrong, and members of the Duke University MADLAB for helpful feedback, and the Wellcome Trust for funding support (092217/Z/10/Z).

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

Funding

Funding for research reported in this article was provided by the Wellcome Trust (092217/Z/10/Z) and the Wellcome Trust Institutional Strategic Support Fund (H2RZKC00).

Note

[1] This effect depends on the relative weighting of payoffs to the self and others. When payoffs to others are weighted higher than payoffs to the self, selfishness increases when less evidence is required before making a decision.

References

Agapakis, C. (2013, July 27). Feynman on biology [Blog post]. Retrieved from http://blogs.scientificamerican.com/oscillator/feynman-on-biology/

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

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

Buckholtz, J. W. (2015). Social norms, self-control, and the value of antisocial behavior. *Current Opinion in Behavioral Sciences*, *3*, 122–129. doi:10.1016/j.cobeha.2015.03.004

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*, 277–284. doi:10.1093/scan/nsr094

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

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences, USA*, *107*, 17433–17438. doi:10.1073/pnas.1009396107

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, USA*, *111*, 17320–17325. doi:10.1073/pnas.1408988111

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

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*, 273–292. doi:10.1177/1088868313495594

Cushman, F. (2015). From moral concern to moral constraint. *Current Opinion in Behavioral Sciences*, *3*, 58–62. doi:10.1016/j.cobeha.2015.01.006

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*, 1082–1089. doi:10.1111/j.1467-9280.2006.01834.x

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, USA*, *109*, 6479–6483. doi:10.1073/pnas.1118653109

Dayan, P., & Huys, Q. J. M. (2009). Serotonin in affective control. *Annual Review of Neuroscience*, *32*, 95–126. doi:10.1146/annurev.neuro.051508.135607

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 & Affective Neuroscience*, *7*, 743–751. doi:10.1093/scan/nss069

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). *How, whether, why: Causal judgments as counterfactual contrasts*. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Greene, J. D. (2014a). The cognitive neuroscience of moral judgment and decision-making. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The cognitive neurosciences* (5th ed., pp. 1013–1023). Cambridge, MA: MIT Press.

Greene, J. D. (2014b). *Moral tribes: Emotion, reason and the gap between us and them*. London, England: Atlantic Books.

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*, 1092–1095. doi:10.1126/science.1153651

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*, 451–462. doi:10.1016/j.neuron.2015.06.031

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292. doi:10.2307/1914185

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). *Inference of intention and permissibility in moral decision making*. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*, Article 7455. doi:10.1038/ncomms8455

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A common mechanism underlying food choice and social decisions. *PLoS Computational Biology*, *11*(10), e1004371.

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Sciences*, *7*, 230–242. doi:10.1111/tops.12131

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*, 147–186. doi:10.1080/1047840X.2014.877340

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

Masserman, J. H., Wechkin, S., & Terris, W. (1964). "Altruistic" behavior in rhesus monkeys. *The American Journal of Psychiatry*, *121*, 584–585.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*, 143–152. doi:10.1016/j.tics.2006.12.007

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: Norm violations and their enforcement. *Neuron*, *56*, 14–18. doi:10.1016/j.neuron.2007.09.020

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, *72*, 692–697. doi:10.1016/j.neuron.2011.11.001

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922. doi:10.1162/neco.2008.12-06-420

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*, 549–562. doi:10.1038/nrn3776

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. doi:10.1126/science.1082976

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*, 667–677. doi:10.1016/j.neuron.2010.07.020

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. doi:10.1111/j.1467-9280.2008.02091.x

Wright, N. D., Symmonds, M., Fleming, S. M., & Dolan, R. J. (2011). Neural segregation of objective and contextual aspects of fairness. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*, 5244–5252. doi:10.1523/JNEUROSCI.3138-10.2011

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience*, *33*, 1099–1108. doi:10.1523/JNEUROSCI.1642-12.2013