Applications of Computational Models in Moral Cognition Research

Authors: Zhang Yinhua, Li Hong, Wu Yin, Wu Yin

Date: 2020-02-12T00:00:00+00:00

Abstract

Moral cognition focuses on the information processing underlying moral psychology. In recent years, researchers have begun to apply computational models to moral cognition research to explore how moral cognition is implemented in the brain. However, research on computational modeling of moral cognition is currently in its infancy. The application of computational models (drift-diffusion models, utility models, reinforcement learning models, and hierarchical Gaussian filter models) in behavioral and physiological studies of moral cognition has quantified the cognitive processes and neural mechanisms underlying moral decision-making, moral judgment, and moral reasoning. Furthermore, this new advancement has contributed to understanding antisocial behavior and mental disorders. Finally, computational modeling needs to be refined, and future research needs to address its potential issues.

Full Text

The Application of Computational Models in Moral Cognition Research

ZHANG Yinhua; LI Hong; WU Yin* (School of Psychology, Normal College, Shenzhen University; Shenzhen Key Laboratory of Affective and Social Cognitive Science, Shenzhen 518060, China)

Abstract: Moral cognition focuses on the information processing underlying moral psychology. In recent years, researchers have begun applying computational models to moral cognition research to explore how moral cognition is implemented in the brain. However, research on computational modeling of moral cognition remains in its infancy. Computational models—specifically Drift Diffusion Models, Utility Models, Reinforcement Learning Models, and Hierarchical Gaussian Filter Models—applied to behavioral and physiological studies

of moral cognition have quantified the cognitive processes and neural mechanisms underlying moral decision-making, moral judgment, and moral inference. Additionally, these new developments contribute to understanding antisocial behavior and mental disorders. Finally, computational modeling approaches require refinement, and future research must attend to their potential limitations.

Keywords: moral cognition; computational models; moral decision-making; moral judgment; moral inference

Recently, He Jiankui's team announced the "birth of the first gene-edited babies" (see *Reference News*, 2019-01-23). Many condemned He Jiankui' s actions as clear ethical violations, and his moral character was vigorously questioned. The controversy centered on ethical and moral issues. While researchers have conducted extensive studies in moral cognition, the unique cognitive mechanisms for resolving moral problems remain unclear. As computational modeling methods for behavioral data have matured, researchers have begun applying computational models to moral cognition. Computational models quantitatively describe, through mathematical functions, how option features (e.g., cost, benefit, waiting time) are converted into valence, thereby influencing decisions (Brown, 2014; Charpentier & O' Doherty, 2018; Konovalov, Hu, & Ruff, 2018). Recent studies have employed this approach to describe the computation of moral valence—how external features of moral problems (e.g., benefits, harms) are transformed into internal utility, and how this utility guides moral decision-making, judgment, and inference (Hackel & Zaki, 2018; Hutcherson, Bushong, & Rangel, 2015; Siegel, Estrada, Crockett, & Baskin-Sommers, 2019; Siegel, Mathys, Rutledge, & Crockett, 2018; Yu, Siegel, & Crockett, 2019). This article reviews the nature of moral cognition, the application of computational models in this domain, and how they advance our understanding of moral cognitive processes and related neural mechanisms.

1. Moral Cognition

The He Jiankui case involves (1) He Jiankui' s decision to edit babies' genes; (2) readers' judgments about whether his specific choices were moral; and (3) further inferences about his moral character. These correspond to three dimensions of moral cognition—moral decision-making, moral judgment, and moral inference (this classification follows Yu et al. (2019)†; see Yu et al., 2018). Their definitions are as follows: moral decision-making refers to choices that affect others' interests; moral judgment refers to the process of evaluating whether behaviors or mental states (e.g., emotions, attitudes) are moral, sometimes including judgments about whether certain actions should be punished or rewarded; moral inference refers to forming beliefs about an actor' s moral character (e.g., good or evil) based on observations of morally relevant behaviors (Yu et al., 2019). Below, we introduce psychological research on moral cognition through these three dimensions.

1.1 Moral Decision-Making

Moral decision-making involves whether individual choices harm others'interests. People have self-interested tendencies (Gray, 1987) and weigh trade-offs between honesty/dishonesty, fairness/unfairness, and generosity/selfishness. Taking honesty decisions as an example, will people make honest decisions (forgoing extra benefits from dishonesty) or dishonest decisions (gaining extra benefits)? Previous research indicates that compared to honest individuals, dishonest individuals take longer to forgo benefits obtained through dishonest decisions (Greene & Paxton, 2009). This suggests that dishonest individuals may experience greater cognitive demands when relinquishing dishonest gains. Moreover, when people make dishonest choices, they experience psychological and physiological discomfort (Cohn, Fehr, & Maréchal, 2014; Gächter & Schulz, 2016; Gamer, Rill, Vossel, & Gödert, 2006). To reduce this discomfort, individuals decrease immoral behaviors. Furthermore, individuals with dorsolateral prefrontal cortex damage show reduced sensitivity to honesty issues (Zhu et al., 2014), and amygdala activation negatively correlates with individuals' history of dishonest behavior—the degree of reduced amygdala activation during current dishonest decisions predicts increased dishonesty in subsequent decisions (Engelmann & Fehr, 2016; Garrett, Lazzaro, Ariely, & Sharot, 2016). This demonstrates the important roles of dorsolateral prefrontal cortex and amygdala in honest decision-making. In summary, decisions often require weighing material interests against moral values, but when choosing moral decisions, the weight of material interests decreases, and people care more about moral values such as honesty and generosity.

1.2 Moral Judgment

Moral judgment is based on moral decision-making and refers to people' s evaluations of whether decisions or decision-makers should be rewarded or punished. The trolley dilemma is a commonly used paradigm for studying moral judgment—imagine a runaway trolley about to kill five workers on the track. The decision-maker can either do nothing, resulting in five deaths, or pull a switch to divert the trolley to a side track where one worker will die (Kamm, 2015). Based on people' s moral approval of the two choices, Greene (2007) proposed a dual-process model of moral judgment—deontological and utilitarian moral judgments. Supporting the decision-maker doing nothing represents a deontological judgment (in deontological ethics, "do not actively kill" is a moral duty), while supporting sacrificing one to save five represents a utilitarian judgment (in utilitarian ethics, one death is better than five). The former is emotion-driven, fast, and automatic; the latter is cognition-driven, slow, and requires motivational and cognitive resources. Research shows that when empathy is evoked, individuals make deontological moral judgments more frequently, while when individuals have less contact with victims or tend toward rational thinking, they make utilitarian moral judgments more frequently (Elqayam, Wilkinson, Thompson, Over, & Evans, 2017; Greene, 2014). Further findings indicate that serotonin reduces the likelihood of utilitarian judgments by increasing aversion

to harming others (Crockett, Clark, Hauser, & Robbins, 2010). Conversely, individuals with ventromedial prefrontal cortex damage make abnormally high utilitarian judgments (Koenigs et al., 2007), suggesting that ventromedial prefrontal cortex is a key neural substrate of the intuitive, emotional system and is crucial for normal moral judgment. In summary, harm aversion is a prosocial emotion that directly influences moral judgment and moral behavior and has implications for treating antisocial and aggressive behaviors.

1.3 Moral Inference

Moral inference involves inferring unobservable, internal states (e.g., motivations behind others' actions or their moral character) from observable, known phenomena (e.g., others' explicit behaviors). Recent research has focused on evaluating behaviors—individuals identify features that influence their moral reasoning. Studies show that negative behaviors (e.g., theft) represent individuals' moral character better than positive behaviors (e.g., donation) (Eisenegger, Naef, Snozzi, Heinrichs, & Fehr, 2010; Uhlmann, Pizarro, & Diermeier, 2015). Donations may be driven by other motives (e.g., maintaining social status), providing less information for reasoning, whereas theft motives are mostly negative (e.g., self-interest, antisocial tendencies), making it easier to infer the thief' s moral character. This suggests that moral reasoning is influenced by the amount of available information. Additionally, research shows that people typically give negative evaluations to hypocrites (those who condemn immoral behavior while engaging in it themselves) (Jordan, Sommers, Bloom, & Rand, 2017; Levine, Barasch, Rand, Berman, & Small, 2018). However, when hypocrites acknowledge their immoral behavior, avoiding sending false signals to others, people evaluate them less negatively. This indicates that people dislike false moral signaling. Moreover, harmful behaviors (e.g., stealing a dead chicken from a supermarket) are considered more immoral than harmless but impure behaviors (e.g., cooking and eating one' s dead pet dog), but the latter reflects worse moral character (Uhlmann & Zhu, 2014). This suggests that person-centered moral reasoning is often more important than behavioral consequences or moral rule violations. In summary, moral inference is both deliberative and intuitive (Garon, Lavallée, Estay, & Beauchamp, 2018).

2. Computational Models

Computer development and application have accelerated computational modeling research, providing more advanced and rigorous methods for scientific inquiry. Computational models use mathematical functions to link observable variables in experiments (e.g., stimuli, outcomes, or past experiences) to recent behaviors and quantify different algorithmic hypotheses underlying behavior. By fitting experimental data to models, researchers investigate the algorithms behind behavior and use precise mathematical models to better understand behavioral data.

In recent years, computational models have been widely applied in psychological

research, including perception, decision-making, memory, and learning. Jiang, Summerfield, and Egner (2016) combined computational models with behavioral and neuroimaging data to reveal how different feature expectations (and attention) for visual objects interact in driving perceptual decisions and neural representations, showing that visual objects are the unit of visual selection. Simply put, when one feature of a visual object is unexpected, this prediction error propagates to other features, making other features of the object unexpected, and thus the visual object as a whole becomes unexpected. Additionally, people generate decisions from value expectations acquired through experience. Meder et al. (2017) proposed that simultaneously representing a series of dynamically changing value evaluations during decision-making could serve as a flexible selection mechanism that combines experientially acquired value information with other value features, enabling adaptive decisions in changing environments. To better adapt to the environment, individuals may flexibly adjust the value assigned to options based on external circumstances or their own states, thereby forming subjective preferences. Ai et al. (2018) established a mathematical model combining decision-making with the dynamic retrieval process of memory, demonstrating that changes in subjective preferences are related to memory consolidation during sleep. More valuably, researchers have used computational models to explore learning mechanisms in patients with mental disorders (e.g., post-traumatic stress disorder) and physiological damage (e.g., basal ganglia damage), providing strong evidence for treatments to restore normal function (Brown et al., 2018; Zhu, Jiang, Scabini, Scabini, & Hsu, 2019). These studies have important implications for future research in psychology and clinical medicine.

Indeed, moral cognition holds a pivotal position in daily life and psychology. To elucidate the cognitive processes and neural mechanisms of moral decision-making, moral judgment, and moral inference, applying this powerful computational modeling approach to moral cognition is inevitable. Below, we review computational models widely used in moral cognition and other fields—Drift Diffusion Models, Utility Models, Reinforcement Learning Models, and Hierarchical Gaussian Filter Models.

2.1 Drift Diffusion Model

The Drift Diffusion Model (DDM), originally developed by Ratcliff (1978), describes decision-making as a continuous sampling process in which noisy information accumulates from a starting point to a boundary or threshold corresponding to an option, which is then selected (Ratcliff & McKoon, 2008). The formula is as follows:

$$dy(t) = v(\Delta u) \cdot dt + \sigma \cdot dW$$

In the formula, $y(t)$ is the amount of information accumulated at time t; Δu is the difference between the boundaries of the two options; v is the speed of

information accumulation (i.e., drift rate); σ is the Gaussian noise parameter of the Wiener process dW. Additionally, DDM parameters include starting point bias, boundary height, and non-decision time. The drift rate represents preference intensity—the stronger an individual' s preference for an option, the faster information accumulates toward that option. Each option has a boundary representing the amount of information that must be accumulated before a response is made. The accumulation process is noisy; at any moment, information may point to either boundary, but more often points to the correct boundary. Non-decision components include encoding the stimulus (which drives the decision process) and extracting stimulus dimensions from the stimulus or memory that form the basis of the decision. DDM can reflect underlying cognitive processes in different components of the model, such as the speed of information accumulation, boundary height, and duration of non-decision processes (Mormann, Malmaud, Huth, Koch, & Rangel, 2010; Lerche & Voss, 2019; Voss, Rothermund, & Voss, 2004).

Moreover, DDM considers all behavioral data, including the shape and location of reaction time distributions for both correct and incorrect responses (Ratcliff, Smith, Brown, & Mckoon, 2016; Ratcliff, Thapar, & McKoon, 2004).

DDM was originally applied to reaction time studies of basic perceptual and memory tasks, such as single-item recognition and associative recognition tasks (Ratcliff, 1978; Ratcliff et al., 2004), and perceptual tasks (including brightness, letters, attention orientation, etc.) (Ratcliff, Thapar, & McKoon, 2003; Thapar, Ratcliff, & McKoon, 2003; Smith, Ratcliff, & Wolfgang, 2004). In the last ten to fifteen years, DDM has become increasingly important in psychological and neural mechanism studies of decision-making processes, including perceptual decisions, simple motor decisions, and value-based decisions. Gold and Shadlen (2007) reviewed how basic decision formation elements are implemented in the brain, proposing that decision-making is a process of weighing priors, evidence, and values, and described specific mathematical operations corresponding to key decision elements (including deliberation and emotional endorsement). They also revealed a basic mechanism for the speed-accuracy trade-off in perceptual tasks and variable reaction times in simple motor tasks—a decision rule that compares a changing decision variable (accumulating evidence over time) with a fixed criterion. Additionally, Krajbich, Armel, and Rangel (2010) used DDM to quantitatively predict the relationship between gaze patterns and choices.

The results showed that in a simple extension of DDM, fixation points participate in value integration, quantitatively explaining various relationships between fixation points and choices, as well as considerable choice biases. Moreover, Krajbich et al. found a causal relationship between visual fixation processes and value comparison processes. That is, by exogenously manipulating relative fixation time, individuals may develop biases in their choices.

Eikemo, Biele, Willoch, Thomsen, and Leknes (2017) studied the modulation of opioid drugs on value-based decision-making in healthy humans, using DDM to fit accuracy and reaction time data to reveal bidirectional drug effects on

two decision subprocesses as expected. In summary, DDM can describe how individuals use priors, evidence, and values to form decisions, revealing general principles behind various forms of decision-making (e.g., perceptual decisions, simple motor decisions, and value-based decisions).

2.2 Utility Models

DDM is typically used for experimental tasks with only two alternatives (i.e., binary choice) and requires many trials per condition, whereas Utility Models can better explain situations with more options. In economics, utility functions measure preferences related to a set of goods and services. Utility is often associated with happiness and satisfaction, which are difficult to observe directly. Therefore, economists use utility functions to represent these abstract, unobservable variables (Debreu, 1954). Later, utility functions were applied to social decision-making, conveying the value of available options to decision-makers, prompting them to choose the option with the highest value (i.e., utility). The simple formula is as follows (assuming two options):

$$\Delta V = U_A - U_B$$

In the formula, U_A is the utility of option A; U_B is the utility of option B; ΔV is the individual' s subjective value. In each trial, subjects have different preferences for each option, and only when the subject prefers option A over B does the utility amount of A exceed B. Therefore, individuals only choose option A when $\Delta V > 0$. Typically, the softmax function is subsequently used to estimate subjects' choice probabilities.

In social decision-making, utility models are mainly used to explore social or moral preferences. Researchers have combined utility models with functional magnetic resonance imaging to study the neural representation of social value, evaluating their distribution of interests between self and others (Liu et al., 2019; Qu, Météreau, Butera, Villeval, & Dreher, 2019; Zhong, Chark, Hsu, & Chew, 2016). This approach partly deciphers the neural mechanisms representing potential interests of self and others, which is crucial for understanding social decision-making. Additionally, Lopez-Persem, Rigoux, Bourgeois-Gironde, Daunizeau, and Pessiglione (2017) obtained comparable utility functions across different tasks with high predictive accuracy for choices. This indicates that comparable utility functions can explain not only economic choices but also different motivation-driven behaviors. Notably, utility models assume that individual preferences are fixed, because if behavior changes according to price or budget variations, it would be impossible to determine the extent to which behavioral changes result from price/budget changes versus preference changes.

2.3 Reinforcement Learning Models

The aforementioned Drift Diffusion Models and Utility Models are widely applied in decision-making, while Reinforcement Learning Models are powerful tools for solving uncertainty in decision-making and various learning problems, including game-related problems (e.g., Tesauro & Gerald, 1995), bicycle riding (e.g., Randløv & Alstrøm, 1998), and robot control (e.g., Riedmiller, Gabel, Hafner, & Lange, 2009). Many different reinforcement learning algorithms have been developed to solve these problems (Szepesvari, 2010; Sutton & Barto, 1998). Learning agents optimize the likelihood of obtaining future rewards by forming stimulus-outcome associations through trial and error, thereby flexibly selecting behaviors that obtain rewards. This process is called reinforcement learning. The key to reinforcement learning is prediction error—the difference between expected and obtained events—which is then used to update beliefs about events in the environment (Sutton & Barto, 1998). Additionally, the most typical and widely used reinforcement learning model is the Rescorla-Wagner model, which represents learning through prediction error signals, is conceptually simple, and computationally efficient (Rescorla & Wagner, 1972). The Rescorla-Wagner model assumes that at time k, the brain computes and updates the value of behavioral variable Q_k as follows:

$$Q_{k+1} = Q_k + \alpha \cdot \delta_k$$

In the formula, α is the learning rate; δ_k is the prediction error, the difference between the actual reward received at time k and the expected reward; Q_k is the current expectation; Q_{k+1} is the individual' s expectation of future rewards. The goal of the reinforcement learning system is to learn a behavioral strategy that maximizes the cumulative reward value obtained by the individual's chosen actions or behaviors.

Reinforcement learning models explain the distinction between behavior-based and outcome-based value representations, linking them to automatic and controlled processing, and precisely clarifying the contributions of cognitive and emotional mechanisms to these two types of processing. On one hand, model-based reinforcement learning activates brain regions including the amygdala, hippocampus, and orbitofrontal cortex (Andrews-Hanna, Reidler, Sepulcre, Poulin, & Buckner, 2010; Zsuga, Biro, Papp, Tajti, & Gesztelyi, 2016). Specifically, the amygdala and ventral striatum jointly encode stimuli (i.e., events beyond expected outcomes), while the hippocampus and ventral striatum jointly encode context (i.e., outcome contingencies). Additionally, orbitofrontal cortex is driven by the hippocampus and amygdala, integrating reward-related information into a contextual framework. Therefore, orbitofrontal cortex provides information about expected rewards, thereby calculating reward expectations (Wallis, 2007).

On the other hand, model-free reinforcement learning can also activate the ventral striatum (Zsuga et al., 2016). The reward expectation information provided

by orbitofrontal cortex feeds back to the model-free system, and based on the functional connectivity of the ventral striatum, enables the ventral striatum to combine model-based reward information with model-free reward prediction errors to calculate value signals emitted by the ventral striatum. Therefore, model-based and model-free reinforcement learning are not separate but have functional connectivity.

2.4 Hierarchical Gaussian Filter Model

Reinforcement learning models provide powerful explanations for simple learning and decision-making behaviors and their neural foundations. However, in reality, involving many stimuli and actions, these algorithms have low learning efficiency and cannot timely capture the speed of human learning. One reason for this difference is that humans utilize the inherent structure in real-world tasks to simplify learning problems (Gershman & Niv, 2010). Therefore, improving reinforcement learning models is inevitable. Inspired by the pioneering work of Behrens, Woolrich, Walton, and Rushworth (2007), Mathy, Daunizeau, Friston, and Stephan (2011) proposed a Hierarchical Gaussian Filter (HGF) model for individual learning under various forms of uncertainty (e.g., environmental volatility and perceptual uncertainty). This model includes a state hierarchy that evolves over time as Gaussian random walks, with the magnitude of each walk (except the first level) determined by the next higher level of the hierarchy. The coupling between levels is controlled by parameters that encode prior beliefs about higher-order structures in the environment, enabling the model to explain individual differences in learning, including between-subject differences and within-subject differences across time. HGF can process both discrete and continuous states and can explain deterministic and probabilistic relationships between environmental events and perceptual states. It can derive closed-form update equations for posterior expectations of all hidden states controlling contingencies in the environment, making HGF computationally efficient and capable of real-time learning. These update equations are formally similar to the Rescorla-Wagner model, providing a Bayesian analogy for reinforcement learning theory. The structure of the Rescorla-Wagner model is: current expectation = previous expectation + learning rate × prediction error. The update equations of HGF are shown in Figure 1 [Figure 1: see original paper]:

Figure 1. Comparison between the update equations of the Hierarchical Gaussian Filter and the structure of the Rescorla-Wagner model. $\mu(k-1)$ is the previous posterior probability; $\mu(k)$ is the current new posterior probability (for specific parameters, see Mathy et al., 2011).

Mathy et al. (2014) further elaborated how HGF provides a general method for processing uncertainty in perception, extending the hierarchical structure of HGF to arbitrary numbers, and exploring how to adapt to various forms of uncertainty through minimization of variational free energy encoded in update equations. In summary, HGF provides a new foundation for understanding normal and abnormal learning, placing reinforcement learning in a general Bayesian

framework and connecting it to optimal principles in probability theory. It provides a principled, flexible, efficient, and intuitive framework for solving the problem of perceptual uncertainty in actors.

HGF is a learning model characterized by assuming that when individuals engage in social learning, the process of forming impressions about others occurs at multiple cognitive levels. Here, we use two cognitive levels—explicit and implicit—as examples. The explicit, observable level is others' specific behaviors, while the implicit (hidden) level is the observer' s internal impression of others. HGF can calculate how information at the explicit level (i.e., each observation of others' specific behaviors) drives changes in representations at the implicit level, providing a "generative model." Siegel et al. (2018) chose HGF to explore the computational basis and temporal dynamics of individual moral inference precisely because it can explain the relationship between implicit impressions and explicit observed behaviors, illustrating how explicit behavioral observations drive impression formation. In summary, practical methods are used to develop computational models of human cognition based on reliable probability principles that can explain the richness and complexity of everyday thinking, reasoning, and learning.

3. Application of Computational Models in Moral Cognition

Computational models can estimate implicit, unobservable latent components in moral cognition processes (parameters reflecting cognitive processes). Researchers can explain and predict the specific cognitive processing of these latent components, developing and refining psychological theories of moral cognition. Computational models can connect moral cognition and moral neuroscience, more comprehensively explaining and predicting neural mechanisms of moral cognition through computational models at different levels. For example, researchers use computational models combined with neuroimaging to reveal potential neural activity processes and cognitive processing components that cannot be directly observed in psychological theories but are related to behavior, such as reward prediction error—a key variable in reinforcement learning (Sven, Pauli, Peter, & John, 2017). This section reviews how the Drift Diffusion Models, Utility Models, Reinforcement Learning Models, and Hierarchical Gaussian Filter Models introduced above are applied to moral cognition.

3.1 Application of Computational Models in Moral Decision-Making

When facing choices with different values, people do not always follow the principle of maximizing benefits by choosing higher-value options (Behrens, Hunt, & Rushworth, 2019; Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Crockett et al., 2015). Some studies indicate that the degree to which people consider others' interests and make choices that deviate from their own benefit maximization positively correlates with their moral behavior (Hutcherson et al.,

2015; Yu et al., 2019).

Hutcherson et al. (2015) had subjects decide whether to accept money allocation schemes for themselves and another party to explore generous decision-making. In DDM, each trial' s choice is based on a dynamically changing random relative decision value signal—estimating expectations for allocation schemes compared to default options. When the random relative decision value signal exceeds a threshold, subjects respond (if positive, accept the allocation scheme; otherwise, reject it), with reaction time equaling the sum of information accumulation time and non-decision time. Results showed that generosity toward others negatively correlated with self-weight and starting threshold, and positively correlated with drift rate (Hutcherson et al., 2015; Konovalov & Krajbich, 2019). Additionally, generous errors (mistakenly choosing to give others more money) occurred significantly more frequently than selfish errors (mistakenly choosing to keep more money), indicating that when individuals value their own rewards more than others' , their generous behavior may reflect noise interference rather than genuine prosocial preferences. At the neural level, ventromedial prefrontal cortex and ventral striatum showed stronger activation when processing self-interest, while ventromedial prefrontal cortex, right temporoparietal junction, and precuneus showed stronger activation when processing others' interests. This suggests that processing self-interest and others' interests is represented independently in the brain. Moreover, ventromedial prefrontal cortex combines self-interest and others' interests into an overall value and integrates the total amount of the allocation scheme through DDM algorithms to make choices. By fitting parameters of DDM components—random relative decision value signal, drift rate, boundary height, starting point bias, and non-decision time—it was deduced and tested that compared to selfish decisions, brain regions related to option information accumulation and value calculation were more active before making generous decisions. These findings reveal the neurocomputational mechanisms underlying moral value representation and suggest that prosocial behavior may be promoted by modulating moral value representation in ventromedial prefrontal cortex.

Krajbich, Hare, Bartling, Morishima, and Fehr (2015) used DDM to find that the speed and consistency of social decisions (selfish or generous) could be predicted by model parameters obtained from non-social decisions (e.g., food choices), indicating that decision-making in these two domains may share the same processing pattern. Additionally, regarding whether social decisions involve a single comparison process or dual processes (intuitive and deliberative), Chen and Krajbich (2018) proposed that behaviors attributed to intuition can serve as starting point bias in DDM processes, similar to prior bias in a Bayesian framework. In dictator game tasks, subjects made binary decisions about how to allocate money between themselves and another party. Results showed that under time pressure, prosocial individuals became more prosocial, while under time delay, the number of prosocial individuals decreased. These findings help unify debates about the cognitive processing of social decisions.

Crockett et al. (2014) had subjects decide whether to administer electric shocks to themselves and others in exchange for benefits (the amount of money increased with the number of shocks) to explore moral decision-making. Crockett et al. used Utility Models with parameters including monetary differences and shock differences between options and default options, loss aversion parameters, and harm aversion parameters to quantify the relative value of pain inflicted on self and others. When the harm aversion parameter equals 0, the decision-maker has minimal harm aversion and will accept any level of shock to increase self-benefit; when the harm aversion parameter approaches 1, the decision-maker has maximal harm aversion and will reduce self-benefit to avoid shocks. Subsequently, the softmax function was used to convert trial-by-trial subjective values into choice probabilities. Results showed that even when individuals' decisions were completely anonymous (with no future negative evaluation or punishment), they cared more about others' pain than their own. Moreover, this concern for others' pain was associated with slower responses when making decisions affecting others, consistent with deliberation in moral decision-making. Computational models identified precise boundaries of this prosocial tendency, which is important for understanding human moral decision-making.

Subsequently, Crockett et al. used Utility Models to study physiological and neural mechanisms in moral decision-making. Results showed that increased serotonin levels increased harm aversion and deliberation time during decision-making, while increased dopamine levels had the opposite effect (Crockett et al., 2015). These distinct roles of serotonin and dopamine in modulating moral behavior have important implications for potential treatments of social dysfunction. Individuals with stronger moral preferences showed lower dorsal striatum activation when harming others for benefit, while lateral prefrontal cortex encoded this guilt (Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017). This suggests that harm aversion, as a moral preference, may affect the values guiding our choices. Notably, parameters in Utility Models vary with different moral decision-making problems (e.g., honesty, fairness, generosity) (Gao et al., 2018; Hu et al., 2018; Sáez, Zhu, Set, Kayser, & Hsu, 2015; Strombach et al., 2015; Zhu et al., 2014).

Compared to traditional research methods, Drift Diffusion Models and Utility Models demonstrate the value of computational models and provide new insights into the nature of moral decision-making. Both explain and predict well how the weighting of self-interest and others' interests affects moral decision-making. Compared to informal models, although parameters in Drift Diffusion Models and Utility Models vary with moral decision-making paradigms, researchers have a unified understanding of them, making these computational models more explanatory and facilitating their application in more fields.

3.2 Application of Computational Models in Moral Judgment

In society, because certain behaviors affect other individuals, people judge whether these behaviors are beneficial or harmful to others. Hackel and Zaki

(2018) used an adapted dictator game paradigm where, in each round, donors (high-wealth and low-wealth) chose to share 20% or 50% of donations, while recipients received the shared amount. Recipients were randomly paired with donors (two high-wealth and two low-wealth) and repeatedly chose which donor to interact with. Thus, recipients learned each donor' s generosity (generosity of sharing 20% = 0, generosity of sharing 50% = 1) and reward value (20% or 50% × donation amount). Recipients then completed a reciprocity task, sharing money with each donor in return. Hackel and Zaki used Reinforcement Learning Models to fit recipients' interaction choices, where reward prediction error reflected donors' reward value and generosity. For example, if a donor first shared 20% of the donation and later shared 50%, this would generate a generosity prediction error (i.e., the donor acted more generously than the recipient expected). Recipients reciprocated more to generous donors (Nowak & Sigmund, 2005) because they made a positive moral judgment of the donor, choosing to reward them, thereby reinforcing their own donating behavior. After reinforcement learning, people not only like generous social partners but also those who provide large material rewards (Feldmanhall, Otto, & Phelps, 2018; Hackel, Doll, & Amodio, 2015; Hackel & Zaki, 2018). Thus, moral judgment can be learned dynamically, prompting subsequent researchers to deeply explore the learning processes of moral judgment and prosocial behavior.

Yu et al. (2019) proposed a harm aversion-centered computational model based on Utility Models and Reinforcement Learning Models, unifying research questions on moral decision-making, judgment, and inference, providing unique insights into revealing the mechanisms of moral cognition. Regarding moral judgment, individuals' blame for actors positively correlates with the additional pain caused by choosing to harm others, but negatively correlates with the additional benefits gained. This indicates that although individuals blame actors for harming others' interests, the benefits obtained justify some harm (Crockett et al., 2010; Xie, Yu, Zhou, Sedikides, & Vohs, 2014). In summary, in moral judgment, the effects of obtaining benefits and harming others on the degree of blame for actors are opposite. First, people believe that harming others more than harming oneself to gain benefits, or gaining benefits only by harming others, increases blame for immoral actors. Second, individuals' own harm aversion preferences moderate the effects of obtaining benefits and harming others on blame, so those less willing to cause pain in others care more about harm than benefits and make more extreme blame judgments when evaluating whether actors should be blamed or rewarded. In summary, when actors' negative outcomes affect others, harm aversion leads judges to impose more punishment, hoping to reduce actors' harmful behaviors.

In addition to harm aversion, moral judgment also involves evaluating outcomes of different magnitudes and probabilities, such as the number of lives saved and probability of rescue in trolley dilemmas. Shenhav and Greene (2010) had subjects evaluate the moral acceptability of sacrificing one life to save a larger group, where group size and probability of death from inaction were uncertain, and used simple Reinforcement Learning Models to fit the data. Results

showed that ventromedial prefrontal cortex encoded subjective representations of expected value in life-and-death moral judgments, while ventral striatum was particularly sensitive to expected moral value. Similarly, right anterior insula was particularly sensitive to death probability. This indicates that judging complex moral decisions affecting others' life and death relies on adapting neural circuits involved in more basic, self-interested decisions concerning material rewards. Shenhav and Greene (2014) further used model-based and model-free Reinforcement Learning Models to fit data, discovering a key dissociation between the effects of automatic and controlled processing on moral judgment, assisted by different neural structures. Amygdala activation reflected individuals' aversion to and degree of blame for harmful utilitarian behaviors. In this integrated moral judgment, ventromedial prefrontal cortex was preferentially involved in relative utilitarian and affective evaluation processing (Shenhav & Greene, 2014). Functional connectivity between amygdala and ventromedial prefrontal cortex varied with the role of emotional input in tasks, being lowest in pure utilitarian judgments and highest in pure affective judgments (Shenhav & Greene, 2010, 2014). These findings suggest that the amygdala provides affective evaluation of judged behaviors, while ventromedial prefrontal cortex combines this signal with utilitarian evaluation of expected outcomes to produce a deliberated moral judgment result. In summary, researchers exploring the neural basis of moral cognition have found that brain regions remain consistently activated during moral judgment processes (Crockett et al., 2017; Shenhav & Greene, 2010). Furthermore, computational models can precisely specify the computations provided by brain regions during moral judgment, promoting the development of moral neuroscience and strengthening the link between observed brain and behavioral changes.

3.3 Application of Computational Models in Moral Inference

Moral inference is a broad concept referring to the process by which individuals identify behavioral features that influence their moral evaluations (e.g., behavioral outcomes and actors' intentions), not necessarily reasoning about good and evil. All inferences about others' characteristics through social learning (e.g., person perception and impression formation) can be considered moral inference (Feldmanhall, Dunsmoor, et al., 2018; Hackel et al., 2015; Joiner, Piva, Turrin, & Chang, 2017; Suzuki et al., 2012). In social interaction, inferring others' intentions is a fundamental problem in forming moral impressions. A basic challenge in moral inference is how humans learn about others' characteristics to predict their own decision-making behaviors. Research shows that perpetrators' apologies not only reduce victims' reactive aggression but also change perpetrators' implicit attitudes toward offenders (Beyens, Yu, Han, Zhang, & Zhou, 2015). Therefore, the morality of a behavior largely depends on the actor' s intention, and inferring the intention behind others' behavior is an important component of moral judgment and moral inference.

Siegel et al. (2018) used Hierarchical Gaussian Filter to explore the compu-

tational basis and temporal dynamics of individual moral inference. Subjects (normal college students) predicted and observed a series of choices by two actors —whether to administer painful electric shocks to another person in exchange for money—and evaluated their impressions of the actors' moral character and uncertainty. Individuals' beliefs about actors' moral character were represented by probability distributions, where the mean described beliefs about the actor after each trial, and variance described uncertainty about that belief. Belief updates over time were represented as Gaussian random walks, with update magnitude determined by individual differences representing belief fluctuations. Results showed that individuals had more uncertain moral beliefs about immoral actors than moral actors, accompanied by faster learning rates. This mechanism enables individuals to flexibly update beliefs about others and promotes forgiveness when initial negative moral impressions prove inaccurate.

Table 1 Summary of computational model applications in moral cognition research

Model Type	Studies
Drift Diffusion Model	Chen & Krajbich, 2018; Hutcherson et al., 2015; Krajbich et al., 2015
Utility Model	Crockett et al., 2014, 2015; Gao et al., 2018; Hu et al., 2018; Sáez et al., 2015; Strombach et al., 2015; Yu et al., 2019; Zhu et al., 2014
Reinforcement Learning Model	Yu et al., 2019; Hackel et al., 2015; Hackel & Zaki, 2018; Shenhav & Greene, 2010, 2014; Joiner et al., 2017; Suzuki et al., 2012
Hierarchical Gaussian Filter Model	Siegel et al., 2018, 2019

Note: Yu et al. (2019) proposed a harm aversion-centered computational model based on Utility Models and Reinforcement Learning Models, unifying research questions on moral decision-making, judgment, and inference.

Subsequently, Siegel et al. (2019) also used Hierarchical Gaussian Filter to study the impact of violence exposure on harm learning in male inmates. Results showed that individuals exposed to violence formed overall subjective social impressions and transformed these impressions into social decisions, but this impaired their moral inference ability (believing moral actors untrustworthy and immoral actors more trustworthy), leading to more immoral behavior. This occurs because wrongly attributing bad characteristics to good people damages existing relationships and hinders new relationship formation (Johnson, Blumstein, Fowler, & Haselton, 2013). Therefore, accurately inferring others' moral

character is crucial for healthy social functioning. From moral decision-making to moral inference is a social learning process, and exploring its cognitive and neural mechanisms is important for correcting inmates' cognition and training individuals with mental disorders such as autism and depression to adapt to normal social functions.

Suzuki et al. (2012) used Reinforcement Learning Models to demonstrate that individual imitation of others' decisions includes two levels of learning signals. In imitation learning, individuals simultaneously present two different prediction error signals—imitating others' reward prediction errors and behavior prediction errors. When imitating others' decisions, ventromedial prefrontal cortex imitates others' characteristics to generate predictions, while dorsomedial prefrontal cortex and dorsolateral prefrontal cortex assist behavioral changes to improve predictions. Hackel et al. (2015) also used Reinforcement Learning Models to reveal that individuals encode reward and trait information through feedback during learning tasks. In addition to specific reward processing, trait information (e.g., generosity or selfishness) is encoded through feedback and can dominate reward information during decision-making. Both learning methods are related to prediction error signals in ventral striatum. Impressions of others can also be formed through feedback-based instrumental learning (Hackel et al., 2015). For example, when a classmate shares resources with everyone, they may not only receive returns but also be considered generous, trustworthy, and cooperative. Consequently, they are valued in other situations, such as being more willing to cooperate with them. Additionally, Joiner et al. (2017) discussed self-referential and other-referential reward prediction errors, which are related to activation in multiple brain regions (e.g., striatum, anterior cingulate cortex, prefrontal cortex, and temporoparietal junction), effectively using Reinforcement Learning Models to regulate social learning. The application of computational models promotes exploration of neural mechanisms underlying social learning and enhances explanatory power for moral inference.

4. Limitations and Future Directions

Moral and immoral behaviors are ubiquitous in life, but research on their cognitive processes and neural mechanisms remains in its infancy. This article reviewed three dimensions of moral cognition (moral decision-making, judgment, and inference) and several widely used computational models in moral cognition (Drift Diffusion Models, Utility Models, Reinforcement Learning Models, and Hierarchical Gaussian Filter Models), and summarized how these computational models elucidate the cognitive processes and neural mechanisms of moral psychology. Notably, Drift Diffusion Models, Utility Models, Reinforcement Learning Models, and Hierarchical Gaussian Filter Models do not have a one-to-one correspondence with moral decision-making, judgment, and inference. Computational models are more related to data types and experimental designs rather than such correspondences in psychological processes. For example, combining Drift Diffusion Models with Reinforcement Learning Models for application in

moral cognition research could be a direction for future research. Compared to traditional research methods and informal models, computational models accurately describe cognitive processes of moral decision-making, judgment, and inference and their potential neural correlates. Additionally, researchers using computational models to study moral issues help resolve debates about the central role of harm in moral cognition (Schein & Gray, 2015, 2018).

Since this article focuses on moral cognition, it cannot discuss in detail other fields using the above computational models, such as resource allocation (Konovalov et al., 2018) and mental disorders (Chen, Takahashi, Nakagawa, Inoue, & Kusumi, 2015; Rothkirch, Tonn, Köhler, & Sterzer, 2017). Research in these fields also clearly benefits from computational models. Note that the specific models discussed here may not be fully applicable to all types of social behaviors, so different computational methods may need to be developed. This article focused on several widely used computational models in moral cognition and how they are applied, so other models that can explain moral cognition issues were not covered, such as Multinomial Processing Tree Models (which predefine how different processes operate as experimental inputs and behavioral outputs, mainly used for moral dilemma research; Liu, Ding, Peng, & Hu, 2019; Cameron, Payne, Sinnott-Armstrong, Scheffer, & Inzlicht, 2017; Gawronski, Conway, Armstrong, Friesdorf, & Hütter, 2018) and Partially Observable Markov Decision Process Models (a type of Bayesian model mainly used to explore belief learning in social contexts; Khalvati et al., 2019). Currently, no single computational model can provide a clear and unified mechanism for moral cognition. Just as simply providing robots with a set of "if-then" rules to adapt to specific situations is very difficult because robots may find themselves in infinitely many situations, parameters derived from single studies cannot serve as final conclusions about numerical weights applicable to all components of moral cognition.

The previous sections introduced some advantages of using computational models to study moral cognition. Here, we emphasize potential problems associated with this approach. First, using different models to obtain values, beliefs, or choice processes carries certain risks—the choice of model (rather than behavior itself) determines the focus of researchers' studies. For example, models used to explain belief learning or preferences differ, and at least some differences in the processes driving these behaviors reflect the use of different computational models. Further development in moral cognition will require more unified methods for modeling different types of cognition. This issue can be illustrated by trust research, which mainly manifests as a learning problem—because trusting others makes one vulnerable to betrayal, people must resolve conflicts between potential benefits and at least three other concerns: loss aversion, unfairness aversion, and betrayal aversion (Bohnet & Zeckhauser, 2004). Few studies have examined the neurocomputational mechanisms underlying these aversions in trust-distrust decisions, which could be studied using mixed models that assign weights to these different concerns (Nave, Camerer, & McCullough, 2015).

Second, although computational models can promote researchers' understanding and prediction of moral cognition, they provide limited perspectives on underlying cognition, learning, and processes. Some models fit behavior and brain activity largely because they can flexibly adapt to many different patterns of data. Therefore, empirical research should strive to provide evidence that a model' s latent parameters actually reflect processes that can be selectively changed through experimental intervention (Hill et al., 2017). Ultimately, good models are those that can construct research driving moral cognition, similar to classical theories but now with a more quantitative and mechanistic focus. Essentially, all models are wrong, but some are useful and can contribute to moral cognition theory.

Finally, because the model-building process itself is quite diverse and flexible, ensuring that computational models are not misused and abused is also very important. Lee et al. (2019) proposed a technical and practical approach, including preregistering models, providing models and registering them after exploratory model development, conducting detailed evaluation of models, and registering modeling reports to make psychological modeling more transparent, credible, effective, and stable. Building paradigms suitable for computational modeling may require trade-offs between real-world richness and methodological rigor. Identifying a computational model that provides a good match to behavior or brain activity does not guarantee that the identified model is the best or most accurate model (Mars, Shea, Kolling, & Rushworth, 2012). Additionally, an important and often overlooked aspect of cognitive computational modeling is simulating candidate models based on observed data (Palminteri, Wyart, & Koechlin, 2017). Despite these limitations, computational models benefit quantitative measurement of individual differences in moral cognition independent of self-report, which may be less reliable for measuring traits with strong social desirability components. Furthermore, model parameters can serve as intermediate levels or cognitive phenotypes between biology and phenomenology, describing how specific clinical or subclinical mental states affect moral cognition and behavior, such as depression (Chen et al., 2015; Rothkirch et al., 2017), schizophrenia (Valton, Romaniuk, Steele, Lawrie, & Seriès, 2017), and personality disorders (Tyrer, Reed, & Crawford, 2015). Therefore, using computational models not only greatly promotes our understanding of human morality but is also gradually being applied in computational psychiatry and other disorders, hoping to reduce human suffering from disease.

Computational models describing moral decision-making, judgment, and inference represent the first step toward quantifying moral cognition and objectively guiding understanding of the cognitive processes underlying moral behavior. These computational models describe how inputs to moral choices are transformed into outputs in the form of mathematical equations. The advantage of computational models is that they provide a common mathematical language that can be used to compare effect sizes across different moral cognition studies. As more research applies these computational models, researchers can aggregate them to rise to the theoretical level (e.g., describing how to combine compo-

nents of moral decision-making, judgment, and inference to improve or propose new theories for the moral cognition field) and provide experience and help for clinical fields (e.g., computational psychiatry). Currently, research on computational models in moral cognition is just beginning, and relatively few models can capture most aspects of moral cognition, or the richness and complexity of human morality may not be reducible to a manageable set of mathematical equations—problems awaiting researchers' solutions.

Reference News. China' s "gene-edited babies" shock the world! What awaits He Jianku—. 2019-01-23 Retrieved from http://ihl.cankaoxiaoxi.com/2018/1127/2359328.shtml

Liu, Y., Ding, Y., Peng, K., & Hu, C. (2019). Application of multinomial processing tree models in social psychology. *Psychological Science*, 42(2), 422-429.

References

Ai, S.Z., Yin, Y. L., Chen, Y., Wang, C., Sun, Y., Tang, X. D., ···Shi, J. (2018). Promoting subjective preferences in simple economic choices during nap. *eLife*, 7, e40583.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain' s default network. *Neuron*, 65, 550-562.

Behrens, T. E., Hunt, L. T., & Rushworth, M. F. (2019). The computation of social behavior. *Science*, 324(5931), 1160-1164.

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(9), 1214-1221.

Beyens, U., Yu, H., Han, T., Zhang, L., & Zhou, X. (2015). The strength of a remorseful heart: Psychological and neural basis of how apology emolliates reactive aggression and promotes forgiveness. *Frontiers in psychology*, 6(1611), 1-16.

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4), 467-484.

Brown, J. W. (2014). The tale of the neuroscientists and the computer: Why mechanistic theory matters. *Frontiers in neuroscience*, 8(349), 1-3.

Brown, V. M., Zhu, L., Wang, J. M., Frueh, B. C., King-Casas B., & Chiu, P. H. (2018). Associability-modulated loss learning is increased in posttraumatic stress disorder. *eLife*, 7, e30150.

Cameron, C. D., Payne, B. K., Sinnott-Armstrong, W., Scheffer, J. A., & Inzlicht, M. (2017). Implicit moral evaluations: A multinomial modeling approach. *Cognition*, 158, 224-241.

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

Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature communications*, 9(3557), 1–10.

Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., & Kusumi, I. (2015). Reinforcement learning in depression: A review of computational research. *Neuroscience & Biobehavioral Reviews*, 55, 247–267.

Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516, 86–89.

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

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, 111(48), 17320–17325.

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

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

Debreu, G. (1954). Representation of a preference ordering by a numerical function. *Decision processes*, 3, 159–165.

Eikemo, M., Biele, G., Willoch, F., Thomsen, L., & Leknes, S. (2017). Opioid modulation of value-based decision-making in healthy humans. *Neuropsychopharmacology*, 42(9), 1833–1840.

Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., & Fehr, E. (2010). Prejudice and truth about the effect of testosterone on human bargaining behaviour. *Nature*, 463(7279), 356–359.

Elqayam, S., Wilkinson, M. R., Thompson, V. A., Over, D. E., & Evans, J. S. B. (2017). Utilitarian moral judgment exclusively coheres with inference from is to ought. *Frontiers in psychology*, 8(1042),1–18.

Engelmann, J. B., & Fehr, E. (2016). The slippery slope of dishonesty. *Nature neuroscience*, 19(12), 1543–1544.

Feldmanhall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America*, 115(7), E1690-E1697.

Feldmanhall, O., Otto, A. R., & Phelps, E. A. (2018). Learning moral values: Another' s desire to punish enhances one' s own punitive behavior. *Journal of Experimental Psychology General*, 147(8),1211-1224.

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595), 496-499.

Gamer, M., Rill, H. G., Vossel, G., & Gödert, H. W. (2006). Psychophysiological and vocal measures in the detection of guilty knowledge. *International Journal of Psychophysiology*, 60(1), 76-87.

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., & Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous-and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences*, 115(33), E7680-7689.

Garon, M., Lavallée, M. M., Estay, E. V., & Beauchamp, M. H. (2018). Visual encoding of social cues predicts sociomoral reasoning. *PloS one*, 13(7), e0201099.

Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*, 19(12), 1727-1732.

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2018). Effects of Incidental Emotions on Moral Dilemma Judgments: An Analysis Using the CNI Model. *Emotion*, 18(7), 989-1008.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, 20(2), 251-256.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535-574.

Gray, J. (1987). The economic approach to human behavior: Its prospects and limitations. In Radnitzky, G., Bernholz, P. (Eds.), *The Economic Method Applied Outside the Field of Economics* (pp. 33-49). New York: Paragon House Publishers.

Greene, J. D. (2007). Why are vmPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322-323.

Greene, J. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

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

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

Hackel, L. M., & Zaki, J. (2018). Propagation of economic inequality through reciprocity and reputation. *Psychological science*, 29(4), 454–464.

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O' Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*,20(8),1142–1149.

Hu, Y., He, L., Zhang, L., Wölk, T., Dreher, J. C., & Weber, B. (2018). Spreading inequality: neural computations underlying paying-it-forward reciprocity. *Social cognitive and affective neuroscience*, 13(6), 578–589.

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

Jiang, J. F., Summerfield, C., & Egner, T. (2016). Visual prediction error spreads across object features in human visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 36(50), 12746–12763.

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *npj Science of Learning*, 2(1), 8, 1–9.

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

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

Kamm, F. M. (2015). *The trolley problem mysteries*. Oxford University Press.

Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., Sestito, M., Dreher, J. C., & Rao, R. P. (2019). Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances*, 5(11), eaax8783.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908–911.

Konovalov, A., Hu, J., & Ruff, C. C. (2018). Neurocomputational approaches to social behavior. *Current Opinion in Psychology*, 24, 1–6.

Konovalov, A., & Krajbich, I. (2019). Revealed indifference: Using response times to infer preferences. *Judgment and Decision Making*, 14(4), 381–394.

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.

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.

Lee, M. D., Criss, A., Devezer, B., Donkin, C., Etz, A., Leite, F. P., . . . Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, 2, 141–153.

Lerche, V., & Voss, A. (2019). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychological Research*, 83(6),1194–1209.

Levine, E. E., Barasch, A., Rand, D. G., Berman, J. Z., & Small, D. A. (2018). Signaling emotion and reason in cooperation. *Journal of Experiment Psychology: General*, 147(5), 702–719.

Liu, Y., Li, S., Lin, W., Li, W., Yan, X., Wang, X., ···& Ma, Y. (2019). Oxytocin modulates social value representations in the amygdala. *Nature neuroscience*, 22(4), 633–644.

Lopez-Persem, A., Rigoux, L., Bourgeois-Gironde, S., Daunizeau, J., & Pessiglione, M. (2017). Choose, rate or squeeze: comparison of economic value functions elicited by different behavioral tasks. *PLoS Computational Biology*, 13(11), e1005848.

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

Mathys, C., Daunizeau, J., Friston, K.J., & Stephan, K.E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5(39), 1–20.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K.E. (2014). Uncertainty in perception and the hierarchical gaussian filter. *Frontiers in Human Neuroscience*, 8(825), 1–24.

Meder, D., Kolling, N., Verhagen, L., Wittmann, M. K., Scholl, J., Madsen K. H., ···Rushworth, M. F. S. (2017). Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nature Communications*, 8(1942), 1–12.

Mormann, M. M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, 5(6), 437–444.

Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science*, 10(6), 772–789.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063), 1291–1298.

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

Qu, C., Météreau, E., Butera, L., Villeval, M. C., & Dreher, J. C. (2019). Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. *PLoS biology*, 17(6), e3000283.

Randløv, J. & Alstrøm, P. (1998). Learning to drive a bicycle using reinforcement learning and shaping. Paper presented at the Proceedings of the Fifteenth International Conference on Machine Learning (USA), Madison, Wisconsin (pp. 463–471). The International Machine Learning Society.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural computation*, 20(4), 873–922.

Ratcliff, R., Smith, P. L., Brown, S. D., & Mckoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4), 260–281.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, 50(4), 408–424.

Ratcliff, R., Thapar, A., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics*, 65(4), 523–535.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement. In Black, A. H., Prokasy, W. F. (Eds.), *Current Research and Theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Riedmiller, M., Gabel, T., Hafner, R., & Lange, S. (2009). Reinforcement learning for robot soccer. *Autonomous Robots*, 27(1), 55–73.

Rothkirch, M., Tonn, J., Köhler, S., & Sterzer, P. (2017). Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain*, 140(4), 1147–1157.

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

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

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

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

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, 34(13), 4741–4749.

Siegel, J. Z., Estrada, S., Crockett, M. J., & Baskin-Sommers, A. (2019). Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males. *Nature Communications*, 10(1942), 1–9.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, 2(10), 750–756.

Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision Research*, 44(12), 1297–1320.

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(5), 1619–1624.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., ···Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron*, 74(6), 1125–1137.

Sven, C., Pauli, W. M., Peter, B., & John, O. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife*, 6, e29718.

Szepesvari, C. (2010). Algorithms for reinforcement learning. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4.1, 1–103.

Tesauro, & Gerald. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3), 58–68.

Thapar, A., Ratcliff, R., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychol Aging*, 18(3), 415–429.

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

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

Uhlmann, E. L., & Zhu, L. (2014). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science*, 5(3), 279–285.

Valton, V., Romaniuk, L., Steele, J. D., Lawrie, S., & Seriès, P. (2017). Comprehensive review: Computational modelling of schizophrenia. *Neuroscience & Biobehavioral Reviews*, 83, 631–646.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, 32(7), 1206–1220.

Wallis, J. D. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annual Review of Neuroscience*, 30, 31–56.

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

Yu, H., Siegel, J.Z., Crockett, M.J. (2019). Modeling morality in 3-D: Decision-making, judgment, and inference. *Topics in Cognitive Science*,11(2), 409–432.

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage*, 129, 95–104.

Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., . . . & Hsu M. (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience*, 17 (10), 1319–1321.

Zhu, L. S., Jiang, Y. M., Scabini, D., Scabini, R. T., & Hsu, M. (2019). Patients with basal ganglia damage show preserved learning in an economic game. *Nature Communications*, 10(802), 1–10.

Zsuga, J., Biro, K., Papp, C., Tajti, G., & Gesztelyi, R. (2016). The "proactive" model of learning: Integrative framework for model- free and model-based reinforcement learning utilizing the associative learning-based proactive brain concept. *Behavioral neuroscience*, 130(1), 6–18.

Note: Figure translations are in progress. See original paper for figures.