Neurocomputational mechanisms engaged in moral choices and moral learning

Chen Qu, Julien Bénistant, Jean-Claude Dreher

Neuroscience and Biobehavioral Reviews xxx (xxxx) 1–11

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

Neurocomputational mechanisms engaged in moral choices and moral learning

Chen Qu [a, bc, 1], Julien Bénistant [c, d, 1], Jean-Claude Dreher [c, d, *]

[a] Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education, China
[b] School of Psychology, Center for Studies of Psychological Application, and Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, China
[c] Laboratory of Neuroeconomics, Institut des Sciences Cognitives Marc Jeannerod, CNRS, Lyon, France
[d] Université Claude Bernard Lyon 1, Lyon, France

ARTICLE INFO

Keywords:
Moral decisions
Moral learning
Neurocomputational models

ABSTRACT

The neural circuitry involved in moral decisions has been studied since the early days of cognitive neuroscience, mainly using moral dilemma. However, the neurocomputational mechanisms describing how the human brain makes moral decisions and learns in various moral contexts are only starting to be established. Here we review recent results from an emerging field using model-based fMRI, which describes moral choices at a mechanistic level. These findings unify the field of moral decision making, extend a conceptual framework previously developed for value-based decision making and characterize how moral processes are computed in the brain. Moral dilemma can be modeled as value-based decisions that weigh self-interests against moral costs/harm to others and different types of prediction errors can be distinguished in different aspects of moral learning. These key computational signals help to describe moral choices and moral learning at an algorithmic level and to reveal how these cognitive operations are implemented in the brain. This researches provide a foundation to account for the neurocomputational mechanisms underlying moral decision making.

1. Introduction

Morality is considered to be the set of customs and values that are embraced by a cultural group to guide social conduct (Moll et al., 2005; Decety and Wheatley, 2015). It is the product of evolutionary pressures that have shaped social and motivational mechanisms into uniquely human forms of experience and behavior. Moral cognition can be considered as a subset of social cognition that focuses on the study of behavior involving moral values, which are rules that define what is good or bad within a society. Moral cognition shares with social cognition the description of processes engaged in the representations of others' mental states, personal goals and social norms (Van Bavel et al., 2015). However, moral cognition specifically focusses on behaviors that are both formally and informally encouraged or discouraged in a given society.

The neural circuitry involved in moral cognition has been studied since the early days of cognitive neuroscience (Greene et al., 2001; Moll et al., 2002; Greene et al., 2004). However, the cognitive neuroscience of morality has only recently been liberated from a simple brain mapping approach that linked brain regions to underspecified moral processes, such as when using simple comparisons between two cognitive conditions. There is a need to understand moral cognition mechanistically using model-based fMRI, which allows characterization of computational processes identified by modeling behavior, and elucidation of where in the brain they are implemented. Assessing the best models to account for specific behavior allows us to define the algorithms at work during moral choices (Krakauer et al., 2017). The confrontation between these algorithms and the brain activity holds the advantage to determine the precise computations at stake in some regions and to understand how the brain decides. This approach has recently proven successful in social neuroscience (Charpentier and O'Doherty, 2018; Konovalov et al., 2018; Suzuki and O'Doherty, 2020), but remains rare in moral neuroscience.

In this review article, we draw on a framework originally proposed to describe value-based decision making, in which choices only depend upon individual's preferences (e.g. food choices, etc.) (Rangel et al., 2008; Sescousse et al., 2013; Frost and McNaughton, 2017; Lopez-Persem et al., 2017). Our review selects moral decision studies which adopted a neurocomputational approach, this framework allows us to

* Corresponding author at: Laboratory of Neuroeconomics, Institut des Sciences Cognitives Marc Jeannerod, CNRS, Lyon, France.
E-mail address: dreher@isc.cnrs.fr (J.-C. Dreher).
[1] Equal contribution.

https://doi.org/10.1016/j.neubiorev.2021.11.023
Received 15 March 2021; Received in revised form 15 November 2021; Accepted 16 November 2021

show that a mechanistic understanding of moral decision making is now possible. We distinguish two fundamental levels of analysis of moral cognition: the algorithmic level (what rules does the brain apply for a particular operation) and the implementation level (how the brain implements that operation) (Lockwood et al., 2020a,b). This distinction is based on the classical proposal that information processing can be described and understood at three levels: the computational or goal of the information-processing system; the algorithmic or the rules that the system applies and the implementational of the system (Marr, 1982). Recent extensions of Marr's framework have been proposed to be applied in social, developmental and evolutionary psychology (Lockwood et al., 2020a,b; van Rooij and Baggio, 2021) and modern applications have been introduced to foster mechanistic understanding of brain-behavior relationships through a pluralistic notion of neuroscience (Krakauer et al., 2017). Here, at the algorithmic level, we identify and describe distinct computational mechanisms engaged, such as the valuation and the learning processes. To do this, we provide an integrative description of the main models used in the field (models of inequity aversion, harm-aversion, honesty and guilt aversion, and models of learning in different moral contexts) (Box 1). We pinpoint computational signals described by these formal models originating from the fields of behavioral economics, neuroeconomics and computer science. These signals help to describe how we make decisions which involve trade-offs between self-benefit and others' harm, how we learn new sets of moral norms, how we learn the moral character of others and how we learn based on moral concerns. At the implementational level, we describe the neural substrates of these computational signals. Our review therefore allows us to decompose the neurocomputational mechanisms that underlie how moral decisions and moral learning are made in various contexts.

2. A theoretical framework for understanding moral decisions and moral learning

One influential framework in the field of neuroeconomics proposed that value-based decision making is intimately linked with learning (Rangel et al., 2008). Here, we argue that a similar framework can be used to describe moral decision-making and the way we learn in different moral contexts (Fig. 1). It proposes that moral choices can be decomposed into five distinct processes (Mas-Colell et al., 1995; Sutton and Barto, 2018). The first process consists of the construction of a representation of the moral decision problem, which entails identifying internal and external states as well as potential courses of moral action. For example, when facing a moral dilemma, such as whether to hurt someone for money, individuals need to represent the possible set of actions (hurting someone for money), the moral principles at stake (e.g., harm-aversion, fairness, etc.) and the preferences of the individual (e.g., the extent of his concern for the other's well-being and his concern for his own benefit). The second process, valuation, consists of attributing a specific value to each action under consideration. Formally, valuation of each moral option under consideration is modelled through a range of utility functions encompassing different moral principles, such as the difference between the individual's earnings and the degree of harm inflicted to another person. These benefits/costs are weighted by parameters representing individuals' preferences. The third process is the selection of one of the actions on the basis of the comparison of their value through the computation of a decision value (DV), which is key to preside moral choices. Such value comparison is central when making value-based (non-moral) decisions between two options (Kable and Glimcher, 2007; Domenech et al., 2017) or among a large set of options (Morris et al., 2021). This step is usually modelled using a softmax function that combines the DV for each moral action and computes a probability to choose each action. An action with higher DV is more likely to be selected. For example, one is more likely to choose to inflict harm to another person if one weighs one's own benefit higher than that of the other person (Box 1). Fourth, after imple-

Box 1.
Models of computational signals engaged in moral choices.

A number of models have been used to describe moral decisions in behavioral economics and decision neuroscience. These models formally quantify the decision values assigned to the different options under consideration in line with moral rules. The Decision value (DV) represents the net value of a specific decision option that is under consideration by an agent. The decision value of an option depends on the costs and benefits, which are integrated by means of a subject-specific value function. The higher the decision value of a given option, the more likely it is to be selected. Indeed, during the action selection process, one is more likely to choose the option that bears the highest decision value.

In addition, reinforcement learning (RL) and Bayesian learning, two fields of research that describe the computational signals needed for learning in different situations mathematically, have been used to account for moral learning. RL is the area of machine learning concerned with how agents take actions in an environment to maximize cumulative rewards. A key computational signal needed to update learning is a Prediction Error (PE) signal representing the discrepancy between the predicted and actual outcome/action. Bayesian learning is a probabilistic approach of learning in which individuals hold probabilistic beliefs over outcomes (e.g. prior) that are updated into a posterior distribution according to the Bayes rule.

Decision value signals for moral choices

Inequity aversion model (Fehr and Schmidt, 1999; Gao et al., 2018)

$$DV_i = \pi_i - \alpha * max\left(\pi_i - \pi_j, 0\right) - \beta * max\left(\pi_j - \pi_i, 0\right) \quad (1)$$

DV_i is the decision value of an individual i who chooses how to allocate resources by minimizing the degree of inequity between herself and another person. Here π_i, stands for the payoff of the decider and π_j for the payoff of the other, j.For a given allocation, if it is advantageous for the decider ($\pi_i > \pi_j$), the difference $\pi_i - \pi_j$ is weighted by a parameter α representing her sensitivity to such inequity. If the allocation is disadvantageous for the decider ($\pi_j > \pi_i$), the difference $\pi_j - \pi_i$ is weighted by a parameter β representing her sensitivity to this type of inequity.

Harm aversion model (Crockett et al., 2017)

$$DV_i = (1-\alpha) * \pi_i - \alpha * S \quad (2)$$

DV_i is the decision value of a decider, i, who has to choose between two allocations of money (π_i) and electric shocks (S). S is the number of shocks which can either be delivered to the decider or to another individual, j. α is a harm-aversion parameter which is different when shocks are for the decider ($\alpha = \alpha_{self}$ and $S = S_i$) or for the other participant ($\alpha = \alpha_{other}$ and $S = S_j$). The difference between α_{other} and α_{self} is defined as the decider's moral preference (this difference is high when someone is more harm averse for another person compared to herself).

Model of immoral behavior benefiting oneself or a charity (Qu et al., 2020)

$$DV_i = \alpha * \pi_i + \beta * \pi_j \quad (3)$$

DV_i is the decision value of a decider, i, who has to choose to accept, or not, an allocation of money for himself or a charity (π_i) and an allocation of money to a bad cause (π_j). α is the weight of the monetary gain for the charity or the decider and β is the weight on the monetary gain for the morally bad cause. These two parameters are different when the gain are for the charity ($\alpha = \alpha_{charity}$ and $\beta = \beta_{charity}$) or for the decider ($\alpha = \alpha_{self}$ and $\beta = \beta_{self}$). An index of moral preference was defined as the difference $(\alpha + \beta)_{charity} - (\alpha + \beta)_{self}$. The higher this index is, the

stronger the preference of participants to weight monetary gain for the charity higher than for themselves when controlling the weights of the moral cost in the two dilemmas, respectively.

Honesty preferences model (Zhu et al., 2014)

$$DV_i = (\alpha - \delta) * \pi_i + (\beta - \delta) * \pi_j \quad (1)$$

DV_i is the utility of an individual, *i* (*a.k.a* the sender) who has to choose between truthful or untruthful information to send to an anonymous receiver, *j*, so that it helps her to choose correctly between two outcomes. π_i is the payoff of the sender and π_j the payoff of the receiver. α is the weight of the monetary gain of the sender and β is the weight on the monetary gain of the receiver. δ represents the cost of deceiving the receiver if the sender chooses to send the untruthful information. This cost depreciates the weighted values of both payoffs π_i and π_j.

Guilt aversion model (Chang et al., 2011; van Baar et al., 2019)

$$DV_i = \alpha * \pi_i - \theta * (E(\pi_j) - \pi_j) \quad (2)$$

DV_i is the utility of an individual, *i*, who has to choose to send back an amount of money to another individual, *j*, who trusted him (*a.k.a* the investor). $\pi_i$is the trustee's payoff and π_j is the investor's payoff. α is the weight of the monetary gain of the trustee, $E(\pi_j)$ is the trustee's expectations about the investor's expected payoff and θ is the guilt-aversion parameter, weighting the difference between the investor's monetary expectation and his actual payoff for a given decision.

Updating signals for learning in different moral contexts

Reinforcement learning model (Lockwood et al., 2016)

$$Q_{t+1}(a) = Q_t(a) + \alpha * (r_t - Q_t(a)) \quad (3)$$

The reinforcement learning model assumes that the associate value of an action *a* is updated as new information is revealed. The action value $Q_{t+1}(a)$ is equal to $Q_t(a)$, the previously estimated action value, plus a prediction error weighted by the learning rate α. The prediction error is the difference between the actual value of the action, r_t and the estimated value $Q_t(a)$. α is the learning rate capturing the individual sensitivity to the prediction error.

The Q-value can be considered as a decision value for the moral choices similar to the ones presented above. Reinforcement learning model, thereby directly linking the valuation/selection processes with the learning process, i.e., the prediction error signal updating the Q-value or Decision value (see Fig. 1).

Bayesian observer (Siegel et al., 2018, 2019)

Individuals form beliefs about other people's harm-aversion tendencies (model 2, Box 1) to be able to infer others' decisions. The probability that another individual chooses one option is defined by individuals' priors on the other's harm-aversion parameters defined by a Gaussian distribution. Beliefs about the distribution of these parameter are updated when individuals observe the actual behavior of the other individual according to Bayes rule:

$$\mu_{i,t} \propto \mu_{i,t-1} + \sigma_{i,t-1} * \delta \quad (4)$$

Here $\mu_{i,t-1}$ is the individuals' prior, updated by the product of the uncertainty over the prior $\sigma_{i,t-1}$ and the prediction error δ. Additionally, a global parameter ω estimates the belief's volatility, which represents the individual's flexibility to update her belief based on the observed behavior of the others.

menting a moral decision, there is a need to evaluate the desirability of the outcome that follows the action. Fifth, the last process, 'learning', consists of updating the representation, valuation and selection processes in order to improve the quality of future moral decisions. Such updates can result from observing others' behavior or from punishment inflicted by others for one's own immoral behavior. The last two steps are often modeled using Reinforcement Learning (RL) or Bayesian learning models, which account for learning of diverse moral behaviors (e.g., altruistic, punitive or trust behavior) (Cushman et al., 2017; Kool et al., 2018; Cushman and Gershman, 2019; FeldmanHall and Dunsmoor, 2019; Cashman and Cushman, 2020) (Box 1). That is, at the time of outcome, a prediction error (PE) signal is computed, that reflects the difference between one's expectation and the actual outcome. This signal is then sent to update representations to be used for future moral decisions.

Using this general framework decomposing how moral choices are computed, we next detail how the two key computations introduced above, decision value and prediction error, can be used to study moral choices and moral learning in distinct contexts. At the algorithmic and implementational levels, we illustrate how this framework accounts for: i) moral choices when weighing personal benefits against the welfare of others (e.g., moral costs such as harm aversion); ii) how we learn new moral norms, how we learn about the moral consequences of our actions, and how we learn about others' moral characters through observation of their actions.

2.1. How do people make trade-offs involving moral principles?

2.1.1. Computations of utility presiding moral choices

Models used to account for moral decisions have been advanced to formalize people's sensitivity to weigh personal benefits against the moral cost of violating internalized moral norms, such as harming others. The moral value of a given action emerges as the integration of moral principles with self and other related information (Van Bavel et al., 2015). As such, moral choices can be accounted for by models of moral preferences developed in the field of behavioral economics. These models, which focus on the valuation and selection processes, propose that the desirability of outcomes expected from alternative options can be quantified by utility functions (Fig. 2a). After attributing a value to each option under consideration, these functions weigh the likely benefits and costs resulting from an action. An option is selected based on maximizing this utility function (Yu et al., 2018). Different models have been used to compute utilities related to moral decisions (Box 1). Such models account for situations in which an individual is facing a dilemma, such as whether to exhibit altruism or reciprocity at a cost to herself (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002). Experimental evidence demonstrates that people consider not only their own material self-interest but also the payoffs of others (Camerer, 2003). This can create dilemmas between one's own pecuniary interests and those of another.

For example, inequity aversion models formalize the computation of the distribution of payoffs for different parties (self and others) into a utility function (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). That is, people can either increase utility from the payoffs of others (Charness and Rabin, 2002) (*i.e.,* the more the others get the higher their own utility) or decrease utility when payoffs are unequal between themselves and others (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000).

Other examples come from harm aversion models and from models of immoral behavior (Box 1). Formally, these models describe how individuals assign decision weights on personal gains and other-regarding components (e.g. moral costs of benefiting from bad actions or harming others). Recent experiments have investigated how people trade off monetary gains/losses against moral costs/benefits (Qu et al., 2019, 2020) or pain to themselves and others (Crockett et al., 2015, 2017). In these types of paradigms, the notions of moral cost and harm aversion are key to explain trade-offs between moral and monetary values (Crockett et al., 2017; Qu et al., 2019, 2020). Moral cost has been implemented as money sent to a morally bad cause and harm aversion as avoiding physical pain (electric shocks) inflicted to a third-party. The models used in these tradeoffs relate specific features of the choice options (e.g., amount of money donated to a bad cause and to oneself, or,

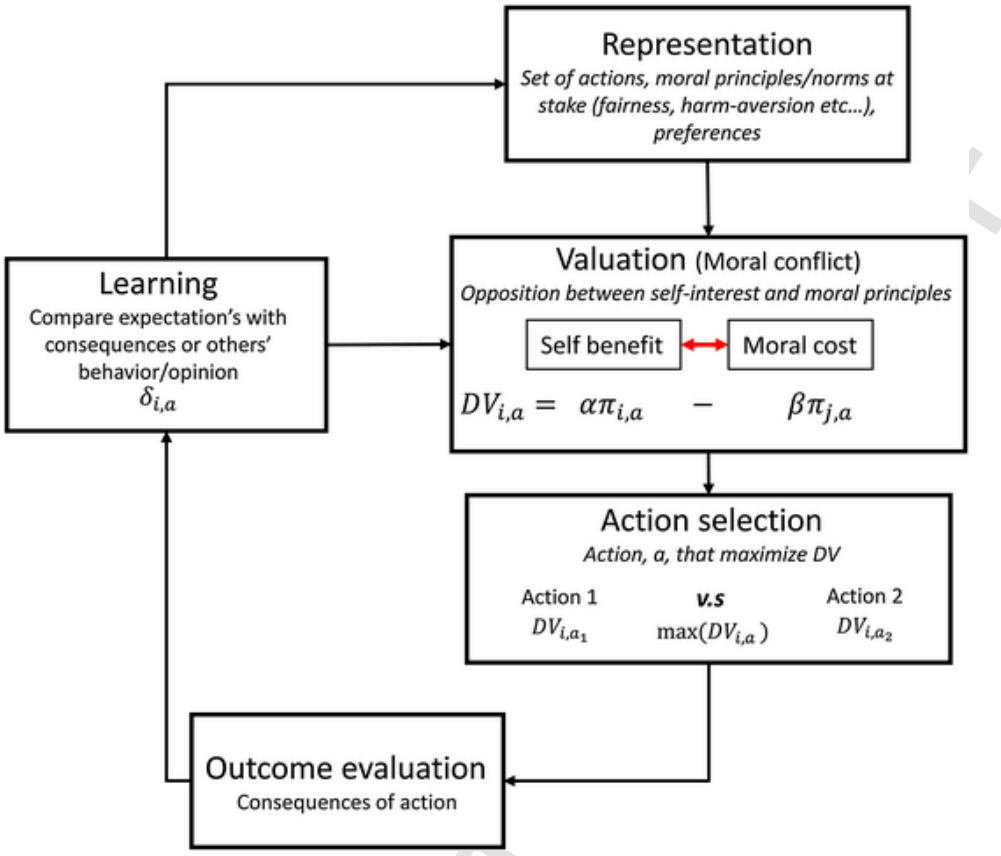

Fig. 1. Conceptual framework describing the computations involved in moral decision-making. The computations involved in moral behavior can be separated into distinct components. The first consists of the representations of the moral dilemma that encompass the moral principles or norms involved as well as internal and external states. Then, individuals evaluate each possible action according to a utility function that weights both the individuals' benefit as well as the consequences for others. The utility function depends on the moral principles at the time of the choice (concerns with fairness, harm-aversion, etc.). Third, individuals select the action that maximizes the Decision Value (DV) computed through the utility function. Fourth, they evaluate the outcome based on the consequences of their moral action and based on other's reaction to it. Finally, a learning signal δ is sent to each of the previous components to update them based on the discrepancy between one's expectations and the observed moral outcome.

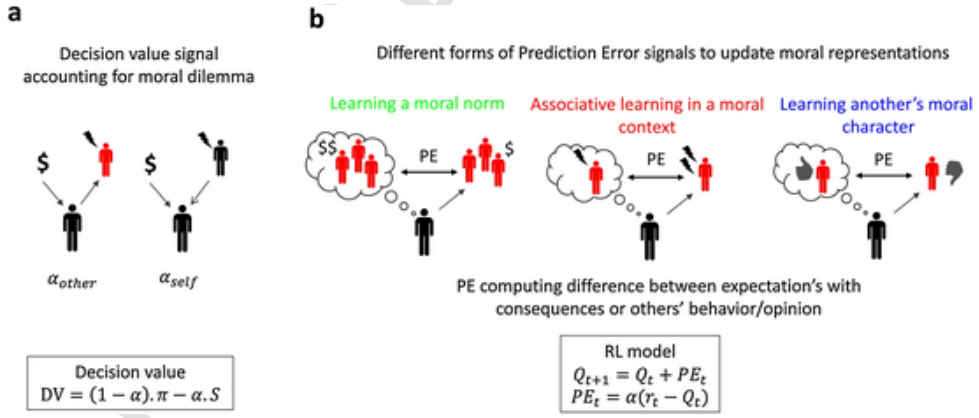

Fig. 2. Computational signals needed for moral decisions and moral learning. a. A key computational signal presiding choice behavior, known as the decision value (DV), weighs the potential benefits and costs of an immoral action (here money is the benefit and harming someone is the cost). Different formulations of DV have been proposed for distinct moral decisions (Box 1). As an illustration, here DV integrates the payoff (π) of an individual (in black) and the shocks she received or she sends to someone else (red person)(S). The parameter α is a harm aversion parameter which differs when oneself or someone else receives the shocks (Crockett et al., 2015). **b.** Illustration of three types of Prediction Error (PE) signals occurring when learning in different moral situations. **Left.** Learning a moral norm is based on the computation of a norm PE that occurs when a deviation from that norm is detected. **Center.** Learning about harmful outcomes for others. A PE occurs when an individual observes an outcome more harmful than expected occurring to another person. **Right.** Learning through observation of the moral actions of others. Learning a person's moral character from observing her actions is achieved via a PE that signals a discrepancy between expectation and observation of her moral behavior.

the difference in the quantity of money and number of electric shocks) to their underlying decision values (Box 1). A softmax function transforms the decision value of accepting offers or choosing harmful options into a probability of making that choice.

In addition to allocation decisions between oneself and others involving tradeoffs between some form of moral cost and monetary benefit, moral decisions also refer to actions that are normally prohibited such as lying, cheating or other dishonest actions. Models of honesty preferences account for the fact that these actions are morally wrong in themselves, independently of their outcomes. Indeed recent theoretical and experimental evidence show that lying is associated with a cost linked to the act itself (Zhu et al., 2014; Kajackaite and Gneezy, 2017) (Box 1).

A last example of a formal model used in moral decision making concerns how guilt and its anticipation arise from morally bad actions. Guilt can be conceptualized as anticipation of a negative emotional state associated with the violation of personal moral rules, social standards or another's expectations (Haidt, 2003; Battigalli and Dufwenberg, 2007). Recent models of guilt-aversion allow individual utility functions to encompass beliefs, a feature essential for modeling emotions. Such formal models provide a precise quantification of the amount of guilt anticipated as the result of a given decision (Gong et al., 2019) (Box 1). According to these models, one's aversion to the possibility of experiencing future guilt prompts morally aligned decisions to minimize guilt anticipation. This approach provides a principled method for elucidating the neural responses to feelings of guilt and exploring how they directly guide moral decision making. Together, these studies show that mathematical models can be used to express moral decisions as tradeoffs between moral values and monetary values.

2.2. How do people learn key representations that guide moral decisions?

2.2.1. Computation of prediction error signals to learn in distinct moral situations

As described in the framework describing moral choices (Fig. 1), there is a bidirectional relationship between moral decision making and learning from PE. For example, moral choices that lead to discrepancy with one's moral values may generate a PE signal. In turn, the outcome of a moral choice that is better than expected generates a positive PE, which will increase the probability of this choice in the future. Similarly, key representations guiding moral decisions, such as harm aversion or norm violation (e.g., being treated unfairly) may be learned via generation of a discrepancy between expectations and outcomes (norm PE), leading to updating of representations for future choices (Kishida et al., 2012; Xiang et al., 2013; Gu et al., 2015; Hétu et al., 2017) (Box 1). Uncertainty regarding one's own moral preferences can also lead the decision making process to be a form of learning whereby one discovers one's preferences by making choices and then observing one's own reactions to those choices (Crockett, 2016). Similarly, uncertainties about outcomes and others' preferences affect moral decision making and learning from others through observing their behavior and their outcomes (Siegel et al., 2018; Khalvati et al., 2019; Park et al., 2019).

2.2.2. Distinguishing different types of moral learning situations

The concept of PE can be applied to different forms of moral learning. Here, we distinguish three types of moral learning phenomenon: (i) learning a new set of moral rules/norms (e.g., when an individual migrates to a new culture/environment); (ii) associative learning based on moral concern (e.g., arbitrary associations between cues and punishments such as electric shocks to an innocent person); (iii) learning the moral character of strangers through observation of their actions (Fig. 2b). There are similarities and differences between these different forms of learning processes. All of them can be formalized by different types of PE: norm PE (Xiang et al., 2013), PE about harmful outcomes occurring to others (Lockwood et al., 2016; Nostro et al., 2020), and PE by observation of other's moral actions (Bellucci et al., 2019; Park et al., 2020). When learning a new set of moral rules and when learning the moral character of strangers, what is learned is moral information. Yet, in the first case, it concerns an environment, in the second it is about an agent. Another difference between learning a new set of moral norms and associative learning based on moral concerns is that in this latter case the moral rule is presumed (e.g., "to avoid harming others") and serves as the underlying motivation that drives learning.

To account for these different types of moral learning, recent psychological theories have used Reinforcement Learning (RL) and Bayesian models (Cushman et al., 2017; Kool et al., 2018; Cushman and Gershman, 2019; FeldmanHall and Dunsmoor, 2019; Cashman and Cushman, 2020) (Box 1). RL models help to explain how a history of pairing social phenomena with positive or negative outcomes can influence and bias complex moral behaviors (Buckholtz, 2015; Gęsiarz and Crockett, 2015; Christopoulos et al., 2017; FeldmanHall et al., 2018a,b; FeldmanHall and Dunsmoor, 2019). For example, RL mechanisms describe learning about others' moral values based on their preference to punish fairness violation (FeldmanHall et al., 2018a,b), or learning others' moral traits, such as generosity (Hackel et al., 2020), honesty (Bellucci et al., 2019) and trustworthiness (Fouragnan et al., 2013; Park et al., 2020) as well as learning moral norms (Xiang et al., 2013; Gu et al., 2015; Hétu et al., 2017). More generally, learning procedures described by Pavlovian and instrumental conditioning provide valuable frameworks for understanding learning in moral contexts, and account for how histories of past decisions influence future moral choice (Gęsiarz and Crockett, 2015). More recently, Bayesian models have been used to account for learning about the moral characters of others (Siegel et al., 2018, 2019) and for the ability to learn moral rules, such as criminal laws and religious commandments (Cushman et al., 2017; Siegel et al., 2018, 2019) (Box 1). This family of models are based on the Bayesian inference mechanism. Individuals start with a probability distribution (prior) over the moral characters of others and update it while observing their actions. This revision of one's prior gives a new probability distribution called *posterior* probability. Over repeated observations, one's prior converge towards the true probability distribution of others' moral characters. Unlike RL models, the Bayesian approach account for the degree of uncertainty of individuals through some parameters such as the variance of the prior distribution. The main result from the literature is that moral inference is explained by an asymmetric Bayesian updating mechanism in which beliefs about the morality of bad agents are more uncertain than beliefs about the morality of good agents. These Bayesian models suggest that negative moral impressions destabilize beliefs about others, promoting cognitive flexibility in the service of cooperative but cautious behavior. One possible extension of Bayesian models in the context of moral decision-making could be to combined a Bayesian learning model with a Bayesian decision model. For example, one previous Bayesian model could be extended to decisions and learning in a moral context (Devaine and Daunizeau, 2017).

Computational models are not only useful to better understand the behavioral processes engaged in moral decisions. These models can also help us to identify the brain areas that support these processes. One key question is to identify the brain regions engaged in the valuation stage at the time of moral choice and the brain areas that integrate the information helpful for this valuation. Another important question is to identify the brain regions computing the PE signals needed to update the moral representations when learning in diverse moral situations.

3. Brain regions engaged in moral valuation and in moral learning

3.1. Brains regions engaged in moral valuation

fMRI combined with models of utility allows the characterization of the brain system that tracks decision value signals when making moral

choices. Neuroimaging research on moral choices has concentrated on cost/benefit tradeoffs such as "Do I ignore my moral values to earn money?" The principle of value computation has proven useful to identify a brain valuation system that includes the ventromedial prefrontal cortex (vmPFC) and ventral striatum. This system is known to be engaged in evaluating primary and secondary rewards, when making social choices (Park et al., 2017; Konovalov et al., 2018; Suzuki and O'Doherty, 2020) and when processing social rewards such as good reputation, being treated fairly, and being cooperative (Rilling et al., 2002; Izuma et al., 2008; Zaki and Mitchell, 2011).

3.1.1. *Do moral value computations engage only the classical brain valuation system?*

Important questions, such as knowing how moral considerations are incorporated into the valuation process, cannot be arbitrated without reference to the brain. Model-based studies have generated three distinct hypotheses regarding how moral considerations may or may not be incorporated in the valuation system. The first hypothesis proposed that computing moral values relies on the same neurocomputational mechanisms as those involved in non-moral value computation. Thus, the brain valuation network classically engaged in value-based decisions would also be engaged for moral decisions during choices coupling financial rewards with moral consequences (Fig. 3a). Supporting this view, several fMRI studies report that the brain has developed the capacity to incorporate moral considerations into its standard valuation circuitry (Hare et al., 2010; Hutcherson et al., 2015; Crockett et al., 2017; Qu et al., 2019, 2020; Hu and Hu, 2021).

A second hypothesis states that, in addition to the classical valuation system, there may also be distinct neural substrates engaged by moral value computation, which preside choices that weigh moral against monetary cost/benefit. According to this account, the computational principles underlying valuation of moral and value-based decisions are similar (weighing self-monetary profits against moral costs/harm). However, moral decisions also engage brain regions not observed in non-moral value-based decision making (Fig. 3a). In one study, the decision value reflecting a trade-off between moral cost and self-monetary benefit engaged the lateral PFC and the anterior insula (Qu et al., 2019). In contrast, a decision value signal encoding the difference between self-monetary cost and compliance with one's moral values (i.e. moral benefit) engaged the ventral putamen (Qu et al., 2019). This is consistent with an early theoretical proposal suggesting that there may be distinct valuation systems for the two types of considerations: one treating violations of moral norms as aversive outcomes, and another treating compliance with moral rules as a rewarding outcome (Rangel et al., 2008). Another recent fMRI study also indicates that moral considerations do not simply engage the standard valuation brain system, since the rTPJ was observed to be specifically engaged in encoding moral values (Ugazio et al., 2019). These findings indicate that similar computational rules are applied by brain systems outside of the classical brain valuation system.

Additional evidence also supports that moral decision computations require nodes outside the classical brain valuation system, including the dlPFC, insula and the rTPJ. For example, the lPFC responds more strongly when harming others for a small relative to a larger profit (Crockett et al., 2017), agreeing with previous work showing that lPFC responds to moral norm violations (Chang and Koban, 2013; Ruff et al., 2013). Altruistic people, who show higher positive moral preference scores, have to overcome a stronger subjective moral costs to accept offers that profit themselves at the expense of their moral values. This behavioral effect is associated with stronger dlPFC signal (Qu et al., 2020). This is also consistent with an association between lPFC responses to immoral earnings and inter-individual differences in other-oriented harm aversion (Crockett et al., 2017) or corruption-related preferences (Hu and Hu, 2021). Neurocomputational studies focusing on dishonesty and guilt aversion also demonstrate a key role of lPFC in computing a variable consistent with moral utility. This included

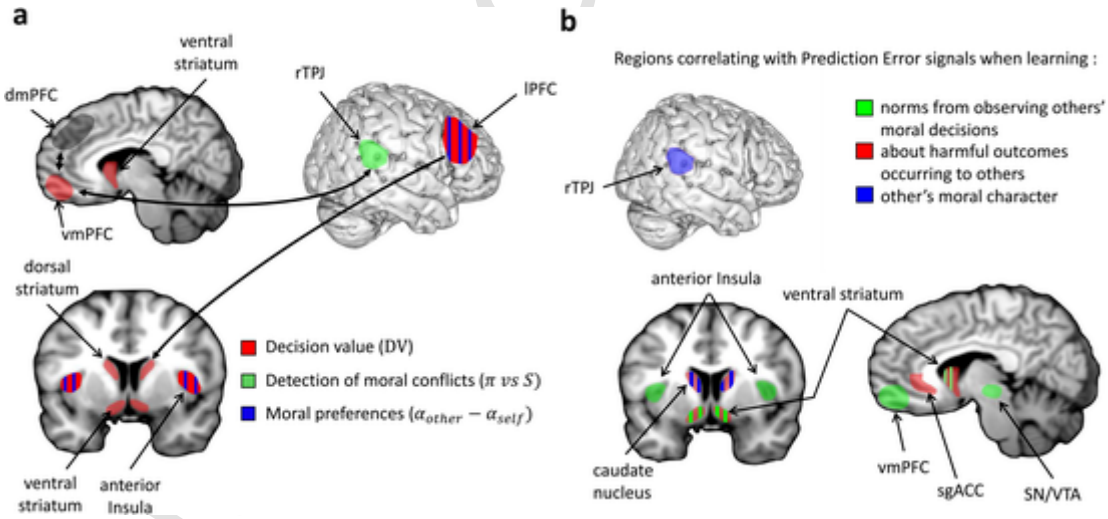

Fig. 3. Brain networks involved in moral decision-making and learning. a. Brain network computing decision value in moral context. This network includes the vmPFC, striatum, lPFC and anterior insula (Hare et al., 2010; Hutcherson et al., 2015; Crockett et al., 2017; Qu et al., 2019, 2020). The rTPJ signals a moral conflict, π *vs* S, reflecting the discrepancy between one's self-interest and moral rules (green) (Obeso et al., 2018)), and the lPFC encodes moral preferences, $\alpha_{other} - \alpha_{self}$, reflecting the individual's degree of adherence to moral rules (blue) (Zhu et al., 2014; Crockett et al., 2017; Gao et al., 2018; Qu et al., 2020). Translating moral norms into moral behavior involves changes in functional connectivity between brain regions, as reflected by a relationship between moral preferences and reduced responses to profiting from others' pain in the dorsal striatum, which is functionally connected with the lPFC (Crockett et al., 2017). Another example comes from the vmPFC, that computes the decision value of an immoral offer, and enhances its functional coupling with components of the mentalizing network (rTPJ and dmPFC), depending upon the beneficiary of an immoral action (Qu et al., 2020). **b. Neural correlates of Prediction Error signals engaged in moral learning.** When learning norms by observing others' moral decisions, PE is encoded in the SN/VTA, ventral striatum, vmPFC and anterior insula (*green*) (Xiang et al., 2013; Hétu et al., 2017). When learning about harmful outcomes occurring to others PE is encoded in the ventral striatum, caudate nucleus and sgACC (*red*) (Lockwood et al., 2016; Lockwood et al., 2021; Zahn et al., 2020). When learning another's moral character PE is encoded in the rTPJ and the caudate nucleus (*blue*) (Fouragnan et al., 2013; Park et al., 2020). *Abbreviations*: rTPJ, right temporo-parietal junction; sgACC, subgenual anterior cingulate cortex; SN/VTA substantia nigra/ventral tegmental area; vmPFC ventromedial prefrontal cortex; mOFC medial orbito frontal cortex; lPFC, lateral prefrontal cortex, dmPFC dorsomedial prefrontal cortex.

weighing differences between behaving honestly and pursuing self-interest (Greene and Paxton, 2009; Zhu et al., 2014; Dogan et al., 2016), advantageous inequity aversion when individuals receive more than others (Nihonsugi et al., 2015; Gao et al., 2018), and maximizing anticipated financial reward while simultaneously minimizing anticipated guilt (Chang et al., 2011; Chang and Koban, 2013; Maréchal et al., 2017; van Baar et al., 2019). Causal evidence for a role of the dlPFC in moral decision value computation was demonstrated in honesty and guilt aversion paradigms performed in patients with focal lesions (Zhu et al., 2014) and in healthy participants using transcranial Direct Current Stimulation (tDCS) (Nihonsugi et al., 2015; Maréchal et al., 2017; Hu and Philippe, 2021). Similarly, the rTPJ is necessary for signaling moral conflicts between self-financial gains and moral values (Obeso et al., 2018). Together, these approaches indicate that neural computations engaged in moral tradeoffs do not simply engage the brain valuation system, but that other areas are necessary to perform moral decision computations. By interfering with a given model-derived signal, they also allow the testing of causal relationships between model components and neural signals (Zhu et al., 2014; Obeso et al., 2018).

A third, non-exclusive, hypothesis is that the brain areas whose activities correlate with utility may not encode utility itself, but attributes of utility (e.g., monetary benefits for oneself, moral cost of hurting others or moral intention). Attributes of moral values may thus be encoded in specific brain regions, and subsequently passed to other areas for integration. Such a view was originally tested for the value attributes of food (e.g., taste, calories, etc…) (Lim et al., 2013; Suzuki et al., 2017). Similarly, in the moral domain, value signals in the vmPFC may integrate inputs from the posterior superior temporal cortex, known to encode attributes such as the intentions of others (Hare et al., 2010), and from the DLPFC, which may act as a domain-general mechanism for representing different attributes in a goal-sensitive manner (Tusche and Hutcherson, 2018).

In addition to the brain system described above, the strength of the relationships between nodes of this network seem to be key to orchestrate moral decisions seamlessly (Fig. 3a). For example, the vmPFC, which computes the decision value of an immoral offer (weighed sum of monetary gain against moral cost), enhanced its functional coupling with nodes of the mentalizing network (rTPJ and dmPFC), depending upon the beneficiary (self vs charity) of an immoral action (Qu et al., 2020). Moral decisions also modulate functional connectivity between the lPFC and the dorsal part of the striatum, which is sensitive to profit, suggesting that moral behavior is linked to a neural devaluation of reward performed by a prefrontal modulation of striatal value representations (Crockett et al., 2017). In addition, the dlPFC and dmPFC are more tightly coupled with the inferior frontal gyrus when the costs of lying are higher in participants who valued honesty highly (Dogan et al., 2016). Similarly, defying a social norm (e.g., acting selfishly when others are generous) increased dlPFC-vmPFC connectivity (Hackel et al., 2020). Together, these studies demonstrate that brain connectivity patterns between nodes of the moral decision brain network depend upon the moral context weighing monetary advantages against moral costs.

Finally, it worth noting that previous studies using hypothetical moral dilemmas and non-computational approaches have pointed out regions such as vmPFC, dlPFC or rTPJ as central in moral cognition (e.g., Greene et al., 2004; Koenigs et al., 2007). They also report others regions such as amygdala whose precise computational role remains to be uncovered. More generally, other value-based decision making studies show that the decision value is sometimes encoded in a wider brain network than the one reported in Fig. 3a (Basten et al., 2010; Bartra et al., 2013; Sescousse et al., 2013). Overall, this advocates for further research and to apply models of moral decision-making to a wider scope of tasks and dilemmas.

3.2. Brain regions encoding updating signals needed for learning in different moral contexts

Model-based fMRI has allowed researchers to make progress in pinpointing the brain regions computing different types of PE engaged in associative learning in moral contexts, in learning the moral characters of others and in learning new sets of moral norms (Fig. 3b).

3.2.1. How does the brain learn about morally-right actions?

Associative learning from PE for outcomes in the moral domain was investigated while learning to avoid electric shocks for either oneself or another person (Lockwood et al., 2020a,b) or learning to choose between options paired with probabilistic monetary rewards for oneself and shocks for a confederate (Nostro et al., 2020). When learning to avoid harm to others *versus* self, a stronger relative balance was observed toward model-free over model-based learning (Lockwood et al., 2020a,b). The caudate nucleus distinguished PE for avoiding harm to others *versus* self (Lockwood et al., 2020a,b). Ventral striatum encoded PE of pain avoidance for self and others and the subgenual anterior cingulate cortex (sgACC), a region known to be implicated in moral agency and responsibility (Zahn et al., 2020), was engaged when deciding to stay vs. switch after no pain for others vs self. The sgACC is also engaged for prosocial PE (i.e gaining rewards for others) (Lockwood et al., 2016; Lockwood and Wittmann, 2018) and in receiving unexpected positive feedback from others (Will et al., 2017). Moreover, individuals with higher empathic concern displayed stronger sgACC activation when deciding to sacrifice their money to prevent others from receiving an electric shock (FeldmanHall et al., 2012, 2015).

3.2.2. How does the brain learn the moral character of others through observation of their actions?

When learning others' moral characters from their actions, a few studies identified a brain system including the rTPJ and caudate nucleus. This network responds with computational variables evolving with the impression of agents' moral character and according to the way they shape subsequent moral judgments (Fouragnan et al., 2013; Bellucci et al., 2019). The rTPJ is associated with a PE signal updating the impressions of others morality (Park et al., 2020). This is also true for the caudate nucleus activity that is associated with updating others' trustworthiness (Fouragnan et al., 2013). Additionally, the dACC computes belief updates during judicial judgments when conforming to other jurors (Park et al., 2017).

3.2.3. How does the brain learn new moral norms?

When individuals learn norms about fairness in a social group, a PE is encoded in brain regions engaged in moral choices, including the vmPFC/mOFC, the anterior insula and the striatum (Xiang et al., 2013). Interestingly, this network displays only a partial overlap with the regions encoding the PE for other types of moral learning. Future work will need to determine whether this finding reflects a truly different neural implementation of similar computational PE principles, or is only due to the scarcity of studies investigating PE signals for different forms of moral learning. Together, these findings illustrate how the use of computational models and the definition of different types of PE are key to understanding the brain systems underlying moral learning.

Many outstanding questions remain to be investigated. In particular, it remains unclear whether the learning processes involved in moral cognition engage distinct algorithms and different neural implementations from non-moral learning. A similar debate has taken place concerning social learning (Ruff and Fehr, 2014; Joiner et al., 2017; Olsson et al., 2020). Some advocate that social learning can be explained by domain-general learning processes (Heyes, 2012, 2018; Heyes and Pearce, 2015; Lind et al., 2019), while others argue that it requires metacognitive knowledge about whom to learn from (Heyes, 2016; Kendal et al., 2018). Future investigations of moral learning will need

to pinpoint their specificity relative to the neurocomputational mechanisms and brain systems engaged with social learning (Apps et al., 2016; Lockwood et al., 2016; Charpentier and O'Doherty, 2018; Konovalov et al., 2018; Khalvati et al., 2019; Basile et al., 2020; Suzuki and O'Doherty, 2020). In particular, it needs to be clarified whether inferring the intentions and motives behind a moral action can be modeled using similar approaches to those proposed for social decisions (Khalvati et al., 2019; Park et al., 2019).

4. Advantages and drawbacks of a neurocomputational approach to moral decisions

Description of the neurocomputational mechanisms engaged in moral cognition provides insights to understand how underspecified cognitive processes can be mapped to computational variables, thereby reducing ambiguity. Mathematical models also allow us to operationalize concepts in a precise fashion and to decompose moral decisions/moral learning into subcomponents (e.g., decision value or prediction error), which are not apparent from direct behavioral observation (Box 1). Computational models also help to predict behavior (Crockett, 2016; Roberts and Hutcherson, 2019) and to test between distinct models reflecting alternative instantiations of different cognitive hypotheses. In addition, computational models can be falsified by showing that a given model is not able to generate a specific behavioral effect of interest (Palminteri et al., 2017). Such approaches help to generate testable predictions and can advance theory by formalizing the components of morality and how they operate at the algorithmic level. This is especially true as some advocate for a more theory-driven study of behavior (van Rooij and Baggio, 2021). A more formal approach of moral decision-making would help to propose testable predictions both at the brain system level and at the behavioral level. For example, the study by van Baar et al. (2019) illustrates how one can use computational models of moral decisions to distinguish various moral strategies and to identify the computational and neural substrates of multiple moral motives underlying reciprocity behavior (e.g., inequity aversion vs guilt aversion). A formal approach of moral choices has also been developed in behavioral economics to study dishonesty and other moral-related decisions (e.g., Gneezy et al., 2018). Another use of formal models lies in their properties to be an oversimplification of the reality. This simplicity allows us to identify the core components needed to explain moral decision making and even failing models can bring valuable insights to better understand these components (Smaldino, 2018). Other advantages of a computational approach to moral choices is that such formal understanding can help to explain how computational variables from different types of moral tasks interact and to generalize results across decision domains (Krajbich and Bartling, 2015; Lopez-Persem et al., 2017; Tusche and Hutcherson, 2018). In particular, formal models help to establish common computational mechanisms between different domains of morality, such as between moral decision-making, judgments of whether an action is morally right or wrong, and moral inferences about others (i.e., "good" vs "bad" people) (Yu et al., 2018). For example, a dynamic model of decision making, initially fitted to participants making food choices was able to predict moral decisions and reaction times of a separate group of subjects (Krajbich and Hare, 2015). Such common computational processes that operate across multiple moral dimensions build bridges across dimensions since individual variability in one dimension of moral cognition may predict variability along other dimensions (Alicke et al., 2015; Uhlmann et al., 2015; Yu et al., 2018).

Cautionary notes about the modeling approach include the fact that any computational model is only a good approximation of behaviors and cognitive processes, within the range of the model space specified. It is important to study inter-individual variability in moral behavior because a single computational model may not perfectly explain the behaviors and cognitive processes of all tested participants. Individual differences may not be best accounted for by the variance in model parameters, there may well be subsets of the population that are better characterized by different models (Moutoussis et al., 2018). It should be acknowledged also that many concepts for moral theories may not be captured by simple computational variables (Roberts and Hutcherson, 2019). Moreover, characterization of the relationships between inter-individual differences in moral preferences and brain activation is still in its infancy (Crockett et al., 2015, 2017; Qu et al., 2020). For example, an elegant study combined computational models with multivariate pattern fMRI analyses to describe inter-individual differences in using different moral strategies (van Baar et al., 2019). Different neural substrates were observed for strategies of guilt aversion, inequity aversion and moral opportunism (in which participants adaptively switch between guilt and inequity aversion strategies). Moral decision related activity patterns in specific brain networks were more similar between participants that shared similar moral strategies for reciprocity decisions than between participants who differed in their strategy. Further work would be needed to specify the underlying neurocomputational mechanisms because these findings only characterized the degree to which specific brain regions selectively process computations relevant to a specific moral strategy.

5. Conclusions

There is a need for understanding moral decision processes at different levels to bridge the gap between fundamental computational principles and the brain system level. In particular, the fact that moral choices and learning rely on computations shared with value-based decision making and social reinforcement learning has been underappreciated. Our proposed framework integrates computational models with model-based fMRI findings to offer a mechanistic explanation for the emergence of moral concerns at the behavioral and neurobiological level. Understanding the computations underlying moral choices may prove useful for the development of AI choice algorithms that concur with the human understanding of morality (eg. in automatic cars) (Awad et al., 2018). Further development of neurocomputational models is also needed to pinpoint the roles of factors such as age, gender, ethnicity, culture, religion, class, and politics on moral cognition (Hester and Gray, 2020; Kelly and O'Connell, 2020), and to move towards more ecological moral scenario commonly encountered (Nastase et al., 2020). Another emerging application concerns computational neuropsychiatry, which would benefit from the elucidation of the dysfunctional neurocomputational mechanisms of different clinical populations engaged in moral behavior (Huys et al., 2016; Lockwood, 2016; Balsters et al., 2017). In particular, autism, psychopathy and major depressive disorder have been described as conditions associated with moral disturbances (Moran et al., 2011; Buon et al., 2013; Fadda et al., 2016; Gong et al., 2019; Schaller et al., 2019; Hu et al., 2020; Zahn et al., 2020). Elucidating the modulating roles of hormones (eg. oxytocin or testosterone) on the neural computations engaged in moral choices may also help to characterize vulnerability to neuropsychiatric diseases (Crockett et al., 2017; Obeso et al., 2018; Li et al., 2020).

Acknowledgments

This research has benefited from the financial support of IDEXLYON from Université de Lyon (project INDEPTH) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) and of the LABEX CORTEX (ANR-11-LABEX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency. This work was also supported by grants from the Agence Nationale pour la Recherche and NSF in the CRCNS program to JCD (ANR n°16-NEUC-0003-01), National Science Foundation of China to QC (31470995). We thank Yang Hu and Edmund Derrington for comments on early versions of the manuscript.

References

Alicke, M.D., et al., 2015. Causal conceptions in social explanation and moral evaluation: a historical tour. Perspect. Psychol. Sci. 10 (6), 790–812. https://doi.org/10.1177/1745691615601888.

Apps, M.A.J., Rushworth, M.F.S., Chang, S.W.C., 2016. The anterior cingulate gyrus and social cognition: tracking the motivation of others. Neuron 90 (4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018PANIST.

Awad, E., et al., 2018. The moral machine experiment. Nature 1. https://doi.org/10.1038/s41586-018-0637-6.

Balsters, J.H., et al., 2017. Disrupted prediction errors index social deficits in autism spectrum disorder. Brain 140 (1), 235–246. https://doi.org/10.1093/brain/aww287.

Bartra, O., McGuire, J.T., Kable, J.W., 2013. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. NeuroImage 76, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063PANIST.

Basile, B.M., et al., 2020. The anterior cingulate cortex is necessary for forming prosocial preferences from vicarious reinforcement in monkeys. PLoS Biol. 18 (6), e3000677. https://doi.org/10.1371/journal.pbio.3000677.

Basten, U., et al., 2010. How the brain integrates costs and benefits during decision making. Proc. Natl. Acad. Sci. 107 (50), 21767–21772. https://doi.org/10.1073/pnas.0908104107.

Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Am. Econ. Rev. 97 (2), 170–176. https://doi.org/10.1257/aer.97.2.170.

Bellucci, G., Molter, F., Park, S.Q., 2019. Neural representations of honesty predict future trust behavior. Nat. Commun. 10 (1), 1–12. https://doi.org/10.1038/s41467-019-13261-8.

Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. Am. Econ. Rev. 90 (1), 166–193. https://doi.org/10.1257/aer.90.1.166.

Buckholtz, J.W., 2015. Social norms, self-control, and the value of antisocial behavior. Curr. Opin. Behav. Sci. 3, 122–129. https://doi.org/10.1016/j.cobeha.2015.03.004PANIST.

Buon, M., et al., 2013. The role of causal and intentional judgments in moral reasoning in individuals with high functioning autism. J. Autism Dev. Disord. 43 (2), 13–458.

Camerer, C.F., 2003. Behavioral game theory: experiments in strategic interaction. Behavioral Game Theory: Experiments in Strategic Interaction. Russell Sage Foundation, New York, NY, US.

Cashman, M., Cushman, F., 2020. Learning from moral failure. In: Lambert, E., Schwenkler, J. (Eds.), Becoming Someone New: Essays on Transformative Experience, Choice, and Change Learning. Oxford University Press.

Chang, L.J., Koban, L., 2013. Modeling emotion and learning of norms in social interactions. J. Neurosci. 33 (18), 7615–7617. https://doi.org/10.1523/jneurosci.0973-13.2013.

Chang, L.J., et al., 2011. Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron 70 (3), 560–572. https://doi.org/10.1016/j.neuron.2011.02.056PANIST.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117 (3), 817–869. https://doi.org/10.1162/003355302760193904.

Charpentier, C.J., O'Doherty, J.P., 2018. The application of computational models to social neuroscience: promises and pitfalls. Soc. Neurosci. 13 (6), 637–647. https://doi.org/10.1080/17470919.2018.1518834.

Christopoulos, G.I., Liu, X.-X., Hong, Y., 2017. Toward an understanding of dynamic moral decision making: model-free and model-based learning. J. Bus. Ethics 144 (4), 699–715. https://doi.org/10.1007/s10551-016-3058-1.

Crockett, M.J., 2016. Computational modeling of moral decisions. In: Forgas, J.P., Jussim, L., Van Lange, P.A.M. (Eds.), The Social Psychology of Morality. pp. 71–88 New York.

Crockett, M.J., et al., 2015. Harm to others outweighs harm to self in moral decision making. Proc. Natl. Acad. Sci. 111 (4), pp. E381–E381. https://doi.org/10.1073/pnas.1424572112.

Crockett, M.J., et al., 2017. Moral transgressions corrupt neural representations of value. Nat. Neurosci. (May), 1–10. https://doi.org/10.1038/nn.4557.

Cushman, F., Gershman, S., 2019. Editors' introduction: computational approaches to social cognition. Top. Cogn. Sci. 11 (2), 281–298. https://doi.org/10.1111/tops.12424.

Cushman, F., Kumar, V., Railton, P., 2017. Moral learning: psychological and philosophical perspectives. Cognition 167, 1–10. https://doi.org/10.1016/j.cognition.2017.06.008PANIST.

Decety, J., Wheatley, T., 2015. The Moral Brain: a Multidisciplinary Perspective. The MIT Press, Cambridge, Massachusetts; London, England, p. 327.

Devaine, M., Daunizeau, J., 2017. Learning about and from others' prudence, impatience or laziness: the computational bases of attitude alignment. PLoS Comput. Biol. 13 (3), 1–28. https://doi.org/10.1371/journal.pcbi.1005422.

Dogan, A., et al., 2016. Prefrontal connections express individual differences in intrinsic resistance to trading off honesty values against economic benefits. Sci. Rep. 6 (1), 33263. https://doi.org/10.1038/srep33263.

Domenech, P., et al., 2017. The neuro-computational architecture of value-based selection in the human brain. Cereb. Cortex, Preprint. https://doi.org/10.1093/cercor/bhw396.

Fadda, R., et al., 2016. Exploring the role of theory of mind in moral judgment: the case of children with autism Spectrum disorder. Front. Psychol. 7, 523. https://doi.org/10.3389/fpsyg.2016.00523.

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. (August), Preprint.

FeldmanHall, O., Dunsmoor, J., 2019. Viewing adaptive social choice through the Lens of associative learning. Perspect. Psychol. Sci. 14 (2), 175–196. https://doi.org/10.1177/1745691618792261.

FeldmanHall, O., et al., 2012. Differential neural circuitry and self-interest in real vs hypothetical moral decisions. Soc. Cogn. Affect. Neurosci. 7 (7), 743–751. https://doi.org/10.1093/scan/nss069.

FeldmanHall, O., et al., 2015. Empathic concern drives costly altruism. NeuroImage 105, 347–356. https://doi.org/10.1016/j.neuroimage.2014.10.043PANIST.

FeldmanHall, O., et al., 2018a. Stimulus generalization as a mechanism for learning to trust. Proc. Natl. Acad. Sci. U.S.A. 115 (7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115.

FeldmanHall, O., Otto, R.A., Phelps, E.A., 2018b. Learning moral values: another's desire to punish enhances one's own punitive behavior. J. Exp. Psychol. Gen. 147 (8), 1211–1224.

Fouragnan, E., et al., 2013. Reputational priors magnify striatal responses to violations of trust. J. Neurosci. 33 (8), 3602–3611. https://doi.org/10.1523/JNEUROSCI.3086-12.2013.

Frost, R., McNaughton, N., 2017. The neural basis of delay discounting: a review and preliminary model. Neurosci. Biobehav. Rev. 79, 48–65. https://doi.org/10.1016/j.neubiorev.2017.04.022PANIST.

Gao, X., et al., 2018. Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. Proc. Natl. Acad. Sci. U.S.A. 115 (33), E7680–E7689. https://doi.org/10.1073/pnas.1802523115.

Gęsiarz, F., Crockett, M.J., 2015. Goal-directed, habitual and pavlovian prosocial behavior. Front. Behav. Neurosci. 9 (MAY), 1–16. https://doi.org/10.3389/fnbeh.2015.00135.

Gneezy, U., Kajackaite, A., Sobel, J., 2018. Lying aversion and the size of the lie. Am. Econ. Rev. 108 (2), pp. 419–153.

Gong, X., et al., 2019. Psychopathic traits are related to diminished guilt aversion and reduced trustworthiness during social decision-making. Sci. Rep. 9 (1), 1–11. https://doi.org/10.1038/s41598-019-43727-0.

Greene, J.D., Paxton, J.M., 2009. Patterns of neural activity associated with honest and dishonest moral decisions. Proc. Natl. Acad. Sci. U.S.A. 106 (30), 12506–12511. https://doi.org/10.1073/pnas.0900152106.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., 2001. An fMRI investigation of emotional engagement in moral judgment. Science 293, 2105–2108.

Greene, J.D., et al., 2004. The neural bases of cognitive conflict and control in moral judgment. Neuron 44 (2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027.

Gu, X., et al., 2015. Necessary, yet dissociable contributions of the insular and ventromedial prefrontal cortices to norm adaptation: computational and lesion evidence in humans. J. Neurosci. 35 (2), 467–473. https://doi.org/10.1523/jneurosci.2906-14.2015.

Hackel, L.M., Wills, J.A., Van Bavel, J.J., 2020. Shifting prosocial intuitions: neurocognitive evidence for a value-based account of group-based cooperation. Soc. Cogn. Affect. Neurosci. 15 (4), 371–381. https://doi.org/10.1093/scan/nsaa055.

Haidt, J., 2003. The moral emotions. Handbook of Affective Sciences. Department of Psychology, University of Virginia, Haidt, Jonathan, pp. 852–870 P.O. Box 400400, Charlottesville, VA, US, 22904-4400: Oxford University Press (Series in affective science.).

Hare, T.A., et al., 2010. Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. J. Neurosci. 30 (2), 583–590. https://doi.org/10.1523/JNEUROSCI.4089-09.2010.

Hester, N., Gray, K., 2020. The moral psychology of raceless, genderless strangers. Perspect. Psychol. Sci. 15 (2), 216–230. https://doi.org/10.1177/1745691619885840.

Hétu, S., et al., 2017. Human substantia nigra and ventral tegmental area involvement in computing social error signals during the ultimatum game. Soc. Cogn. Affect. Neurosci. 12 (12), 1972–1982. https://doi.org/10.1093/scan/nsx097.

Heyes, C., 2012. What's social about social learning? J. Comp. Psychol. 126 (2), 193–202. https://doi.org/10.1037/a0025180.

Heyes, C., 2016. Who knows? Metacognitive social learning strategies. Trends Cogn. Sci. (Regul. Ed.) 20 (3), 204–213. https://doi.org/10.1016/j.tics.2015.12.007PANIST.

Heyes, C., 2018. Enquire within: cultural evolution and cognitive science. Philos. Trans. Biol. Sci. 373 (1743), 20170051. https://doi.org/10.1098/rstb.2017.0051.

Heyes, C., Pearce, J.M., 2015. Not-so-social learning strategies. Proc. R. Soc. B: Biol. Sci. 282 (1802), 20141709. https://doi.org/10.1098/rspb.2014.1709.

Hu, Y., Hu, C., et al., 2021. Neural basis of corruption in power-holders. eLife 10, 1–27. https://doi.org/10.7554/eLife.63922.

Hu, Y., Philippe, R., et al., 2021. Perturbation of right dorsolateral prefrontal cortex (rDLPFC) makes power-holders less resistant to tempting bribes. Psychol. Sci. In Press.

Hu, Y., et al., 2020. Right temporoparietal junction underlies avoidance of moral transgression in Autism Spectrum disorder. J. Neurosci., p. JN-RM-1237-20. https://doi.org/10.1523/JNEUROSCI.1237-20.2020.

Hutcherson, C.A., Bushong, B., Rangel, A., 2015. A neurocomputational model of altruistic choice and its implications. Neuron 87 (2), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031PANIST.

Huys, Q.J.M., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19 (3), 404–413. https://doi.org/10.1038/nn.4238.

Izuma, K., Saito, D.N., Sadato, N., 2008. Processing of social and monetary rewards in the human striatum. Neuron 58 (2), 284–294. https://doi.org/10.1016/j.neuron.2008.03.020.

Joiner, J., et al., 2017. Social learning through prediction error in the brain. NPJ Sci. Learn. 2 (1), 8. https://doi.org/10.1038/s41539-017-0009-2.

Kable, J.W., Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. Nat. Neurosci. 10 (12), 1625–1633. https://doi.org/10.1038/nn2007.

Kajackaite, A., Gneezy, U., 2017. Incentives and cheating. Games Econ. Behav. 102, 433–444. https://doi.org/10.1016/j.geb.2017.01.015PANIST.
Kelly, C., O'Connell, R., 2020. Can neuroscience change the way we view morality? Neuron 108 (4), 604–607. https://doi.org/10.1016/j.neuron.2020.10.024.
Kendal, R.L., et al., 2018. Social learning strategies: bridge-building between fields. Trends Cogn. Sci. 22 (7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003PANIST.
Khalvati, K., et al., 2019. Modeling other minds: bayesian inference explains human choices in group decision-making. Sci. Adv. 5 (11), eaax8783. https://doi.org/10.1126/sciadv.aax8783.
Kishida, K.T., et al., 2012. Implicit signals in small group settings and their impact on the expression of cognitive capacity and associated brain responses. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 367 (1589), 704–716. https://doi.org/10.1098/rstb.2011.0267.
Koenigs, M., et al., 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446 (7138), 908–911. https://doi.org/10.1038/nature05631.
Konovalov, A., Hu, J., Ruff, C.C., 2018. Neurocomputational approaches to social behavior. Curr. Opin. Psychol. 24, 41–47. https://doi.org/10.1016/j.copsyc.2018.04.009PANIST.
Kool, W., Cushman, F.A., Gershman, S.J., 2018. Chapter 7 - competition and cooperation between multiple reinforcement learning systems. In: Morris, R., Bornstein, A., Shenhav, A. (Eds.), Goal-Directed Decision Making. Academic Press, pp. 153–178. https://doi.org/10.1016/B978-0-12-812098-9.00007-3.
Krajbich, I., Bartling, B., et al., 2015. Rethinking slow based on a critique of reaction-time reverse inference. Nat. Commun. 6 (7455), 1–9. https://doi.org/10.3389/fpsyg.2016.01174.
Krajbich, I., Hare, T., et al., 2015. A common mechanism underlying food choice and social decisions. PLoS Comput. Biol. 11 (10), 1–24. https://doi.org/10.1371/journal.pcbi.1004371.
Krakauer, J.W., et al., 2017. Neuron Perspective Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron 93 (3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041.
Li, Y., et al., 2020. Endogenous testosterone is associated with increased striatal response to audience effects during prosocial choices. Psychoneuroendocrinology 122, 104872. https://doi.org/10.1016/j.psyneuen.2020.104872.
Lim, S.-L., O'Doherty, J.P., Rangel, A., 2013. Stimulus value signals in ventromedial PFC reflect the integration of attribute value signals computed in Fusiform Gyrus and posterior superior temporal gyrus. J. Neurosci. 33 (20), 8729–8741. https://doi.org/10.1523/JNEUROSCI.4809-12.2013.
Lind, J., Ghirlanda, S., Enquist, M., 2019. Social learning through associative processes: a computational theory. R. Soc. Open Sci. 6 (3), 181777. https://doi.org/10.1098/rsos.181777.
Lockwood, P.L., 2016. The anatomy of empathy: vicarious experience and disorders of social cognition. Behav. Brain Res. 311, 255–266. https://doi.org/10.1016/j.bbr.2016.05.048PANIST.
Lockwood, P.L., Wittmann, M.K., 2018. Ventral anterior cingulate cortex and social decision-making. Neurosci. Biobehav. Rev. 92 (May), 187–191. https://doi.org/10.1016/j.neubiorev.2018.05.030PANIST.
Lockwood, P.L., et al., 2016. Neurocomputational mechanisms of prosocial learning and links to empathy. Proc. Natl. Acad. Sci. 113 (35), 9763–9768. https://doi.org/10.1073/pnas.1603198113.
Lockwood, P.L., et al., 2020a. Model-free decision making is prioritized when learning to avoid harming others. Proc. Natl. Acad. Sci. 117 (44), 27719. https://doi.org/10.1073/pnas.2010890117.
Lockwood, P.L., Apps, M.A.J., Chang, S.W.C., 2020b. Is there a "Social" brain? Implementations and algorithms. Trends Cogn. Sci. (Regul. Ed.) 24 (10), 802–813. https://doi.org/10.1016/j.tics.2020.06.011.
Lockwood, P.L. et al. (In press) 'Model-free decision-making is prioritized when learning to avoid harming others', Proceedings of the National Academy of Sciences [Preprint]. doi:https://doi.org/10.1101/718106.
Lopez-Persem, A., et al., 2017. Choose, rate or squeeze: comparison of economic value functions elicited by different behavioral tasks. PLoS Comput. Biol. 13 (11), 1–18. https://doi.org/10.1371/journal.pcbi.1005848.
Maréchal, M.A., et al., 2017. Increasing honesty in humans with noninvasive brain stimulation. Proc. Natl. Acad. Sci. 114 (17), 4360–4364. https://doi.org/10.1073/pnas.1614912114.
Marr, D., 1982. Vision: a Computational Investigation Into the Human Representation and Processing of Visual Information. Henry Holt and Co., Inc., USA.
Mas-Colell, A., Whinston, M., Green, J., 1995. Microeconomic Theory Available at:. Oxford University Press. https://EconPapers.repec.org/RePEc:oxp:obooks:9780195102680.
Moll, J., et al., 2002. The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. J. Neurosci. 22 (7), 2730 LP-2736. https://doi.org/10.1523/JNEUROSCI.22-07-02730.2002.
Moll, J., et al., 2005. The neural basis of human moral cognition. Nat. Rev. Neurosci. 6, 799–809.
Moran, J.M., et al., 2011. Impaired theory of mind for moral judgment in high-functioning autism. Proc. Natl. Acad. Sci. U.S.A. 108 (7), 2688–2692. https://doi.org/10.1073/pnas.1011734108.
Morris, A., et al., 2021. Generating options and choosing between them depend on distinct forms of value representation. Psychol. Sci., Preprint. https://doi.org/10.1177/09567976211005702.
Moutoussis, M., Hopkins, A.K., Dolan, R.J., 2018. Hypotheses about the relationship of cognition with psychopathology should be tested by embedding them into empirical priors. Front. Psychol. 9, 2504. https://doi.org/10.3389/fpsyg.2018.02504.
Nastase, S.A., Goldstein, A., Hasson, U., 2020. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. NeuroImage 222, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254.
Nihonsugi, T., Ihara, A., Haruno, M., 2015. Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. J. Neurosci. 35 (8), 3412–3419. https://doi.org/10.1523/JNEUROSCI.3885-14.2015.
Nostro, A.D., et al., 2020. Neuro-computational mechanisms of action-outcome learning under moral conflict. bioRxiv, p. 2020.06.10.143891. https://doi.org/10.1101/2020.06.10.143891.
Obeso, I., et al., 2018. A causal role for right temporo-parietal junction in signaling moral conflict. eLife 7, e40671. https://doi.org/10.7554/eLife.40671.
Olsson, A., Knapska, E., Lindström, B., 2020. The neural and computational systems of social learning. Nat. Rev. Neurosci. 21 (4), 197–212. https://doi.org/10.1038/s41583-020-0276-4.
Palminteri, S., Wyart, V., Koechlin, E., 2017. The importance of falsification in computational cognitive modeling. Trends Cogn. Sci. 21 (6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011PANIST.
Park, S.A., et al., 2017. Integration of individual and social information for decision-making in groups of different sizes. In: Rushworth, M. (Ed.), PLoS Biol. 15 (6), e2001958, Edited by. https://doi.org/10.1371/journal.pbio.2001958.
Park, S.A., et al., 2019. Neural computations underlying strategic social decision-making in groups. Nat. Commun. 10 (1), 5287. https://doi.org/10.1038/s41467-019-12937-5.
Park, B., et al., 2020. The role of right temporo-parietal junction in processing social prediction error across relationship contexts. Soc. Cogn. Affect. Neurosci., [Preprint], (nsaa072). https://doi.org/10.1093/scan/nsaa072.
Qu, C., et al., 2019. Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. PLoS Biol. 17 (6), e3000283, Edited by M.F.S. Rushworth. https://doi.org/10.1371/journal.pbio.3000283.
Qu, C., et al., 2020. Neurocomputational mechanisms underlying immoral decisions benefiting self or others. Soc. Cogn. Affect. Neurosci. (January), 135–149. https://doi.org/10.1093/scan/nsaa029.
Rangel, A., Camerer, C., Montague, P.R., 2008. A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. 9 (7), 545–556. https://doi.org/10.1038/nrn2357.
Rilling, J.K., et al., 2002. A neural basis for social cooperation. Neuron 35 (2), 395–405. https://doi.org/10.1016/S0896-6273(02)00755-9.
Roberts, I.D., Hutcherson, C.A., 2019. Affect and decision making: insights and predictions from computational models. Trends Cogn. Sci. 23 (7), 602–614. https://doi.org/10.1016/j.tics.2019.04.005.
Ruff, C.C., Fehr, E., 2014. The neurobiology of rewards and values in social decision making. Nat. Rev. Neurosci. 15 (8), 549–562. https://doi.org/10.1038/nrn3776.
Ruff, C.C., Ugazio, G., Fehr, E., 2013. Changing social norm compliance with noninvasive brain stimulation. Science 342 (6157), 482–484. https://doi.org/10.1126/science.1241399.
Schaller, U.M., et al., 2019. Intuitive moral reasoning in high-functioning autism Spectrum disorder: a matter of social schemas? J. Autism Dev. Disord. 49 (5), 1807–1824. https://doi.org/10.1007/s10803-018-03869-y.
Sescousse, G., et al., 2013. Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies. Neurosci. Biobehav. Rev. 37 (4), 681–696. https://doi.org/10.1016/j.neubiorev.2013.02.002PANIST.
Siegel, J.Z., et al., 2018. Beliefs about bad people are volatile. Nat. Hum. Behav. 2 (10), 750–756. https://doi.org/10.1038/s41562-018-0425-1.
Siegel, J.Z., et al., 2019. Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males. Nat. Commun. 10 (1). https://doi.org/10.1038/s41467-019-09962-9.
Smaldino, P.E., 2018. Models are stupid, and we need more of them. Computational Social Psychology. pp. 311–331 January 2016 https://doi.org/10.4324/9781315173726-14.
Sutton, R.S., Barto, A.G., 2018. Reinforcement learning: an introduction Cambridge, MA, USA. A Bradford Book.
Suzuki, S., O'Doherty, J.P., 2020. Breaking human social decision making into multiple components and then putting them together again. Cortex 127, 221–230. https://doi.org/10.1016/j.cortex.2020.02.014.
Suzuki, S., Cross, L., O'Doherty, J.P., 2017. Elucidating the underlying components of food valuation in the human orbitofrontal cortex. Nat. Neurosci. 20 (12), 1780–1786. https://doi.org/10.1038/s41593-017-0008-x.
Tusche, A., Hutcherson, C.A., 2018. Cognitive regulation alters social and dietary choice by changing attribute representations in domain-general and domain-specific brain circuits. eLife 7, e31185, Edited by G. Schoenbaum. https://doi.org/10.7554/eLife.31185.
Ugazio, G., et al., 2019. Neuro-computational foundations of moral preferences. bioRxiv 801936. https://doi.org/10.1101/801936.
Uhlmann, E.L., Pizarro, D.A., Diermeier, D., 2015. A person-centered approach to moral judgment. Perspect. Psychol. Sci. 10 (1), 72–81. https://doi.org/10.1177/1745691614556679.
van Baar, J.M., Chang, L.J., Sanfey, A.G., 2019. The computational and neural substrates of moral strategies in social decision-making. Nat. Commun. 10 (1). https://doi.org/10.1038/s41467-019-09161-6.
Van Bavel, J.J., FeldmanHall, O., Mende-Siedlecki, P., 2015. The neuroscience of moral cognition: from dual processes to dynamic systems. Moral. Ethics 6, 167–172. https://doi.org/10.1016/j.copsyc.2015.08.009.
van Rooij, I., Baggio, G., 2021. Theory before the test: how to build high-verisimilitude explanatory theories in psychological science. Perspect. Psychol. Sci. 16 (4), 682–697. https://doi.org/10.1177/1745691620970604.

Will, G.-J., et al., 2017. Neural and computational processes underlying dynamic changes in self-esteem. eLife 6, e28098, Edited by O. FeldmanHall. https://doi.org/10.7554/eLife.28098.

Xiang, T., Lohrenz, T., Montague, P.R., 2013. Computational substrates of norms and their violations during social exchange. J. Neurosci. 33 (3), 1099–1108. https://doi.org/10.1523/jneurosci.1642-12.2013.

Yu, H., Siegel, J.Z., Crockett, M.J., 2018. Modeling morality in 3-D: decision-making, judgment, and inference. Top. Cogn. Sci. 1–24. https://doi.org/10.1111/tops.12382.

Zahn, R., de Oliveira-Souza, R., Moll, J., 2020. Moral motivation and the basal forebrain. Neurosci. Biobehav. Rev. 108, 207–217. https://doi.org/10.1016/j.neubiorev.2019.10.022.

Zaki, J., Mitchell, J.P., 2011. Equitable decision making is associated with neural markers of intrinsic value. Proc. Natl. Acad. Sci. U.S.A. 108 (49), 19761–19766, 2011/11/21 edn. https://doi.org/10.1073/pnas.1112324108.

Zhu, L., et al., 2014. Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. Nat. Neurosci. 17 (10), 1319–1321. https://doi.org/10.1038/nn.3798.