Cognitive, Affective, & Behavioral Neuroscience
2008, 8 (4), 390-401
doi:10.3758/CABN.8.4.390

The role of moral utility in decision making: An interdisciplinary framework

PHILIPPE N. TOBLER
University of Cambridge, Cambridge, England

ANNEMARIE KALIS
Utrecht University, Utrecht, The Netherlands

AND

TOBIAS KALENSCHER
University of Amsterdam, Amsterdam, The Netherlands

What decisions should we make? Moral values, rules, and virtues provide standards for morally acceptable decisions, without prescribing how we should reach them. However, moral theories do assume that we are, at least in principle, capable of making the right decisions. Consequently, an empirical investigation of the methods and resources we use for making moral decisions becomes relevant. We consider theoretical parallels of economic decision theory and moral utilitarianism and suggest that moral decision making may tap into mechanisms and processes that have originally evolved for nonmoral decision making. For example, the computation of reward value occurs through the combination of probability and magnitude; similar computation might also be used for determining utilitarian moral value. Both nonmoral and moral decisions may resort to intuitions and heuristics. Learning mechanisms implicated in the assignment of reward value to stimuli, actions, and outcomes may also enable us to determine moral value and assign it to stimuli, actions, and outcomes. In conclusion, we suggest that moral capabilities can employ and benefit from a variety of nonmoral decision-making and learning mechanisms.

Imagine that you are the democratically elected president of a country and have just been informed that a passenger airplane has been hijacked. The hijackers appear to be headed for a city with well-populated skyscrapers. Your air force is in place and ready to shoot the passenger plane down at your command. You can take many different courses of action: shoot or not, negotiate or not. How should you decide? Would it make a difference whether 10 or 500 passengers are on board the airplane, or whether the passengers have voted for you? Would it make a difference if the air force is not in place and the only way to stop the hijackers in time is to ask the pilot of a nearby second passenger airplane to crash into the first? Would you try to calculate how likely it is that the hijackers actually will crash the airplane into a skyscraper and how many innocent people would die if the hijackers were to do so? Assuming that the probability of the hijackers' heading for a skyscraper appears high, should you have the plane crashed by the other plane with its x additional passengers being killed with certainty, in order to probabilistically save y people on the ground ($y > x$)? Or do not only hijackers, but also presidents, have an unconditional duty not to kill innocent people?

The questions above concern moral decision making, particularly what normative claims should guide us when we make moral decisions and how we actually make such moral decisions. In the present review, we attempt to provide a road map for the empirical investigation of moral decision making. Our starting point will be the important philosophical distinction between *is* and *ought*. This distinction and the need for normative claims sometimes gets forgotten by politicians and members of the public when they make statements about what should be done, given the facts. Then we will briefly introduce utilitarian and deontological moral theories and will review psychological and neuroscientific findings regarding moral decision making. This research will suggest that people may use more than one moral theory, as well as moral heuristics, and that the prefrontal cortex may be particularly important for moral decision making. Finally, we will propose

P. N. Tobler, pnt21@cam.ac.uk

Copyright 2008 Psychonomic Society, Inc.

that moral decisions may rely on mental processes and neuronal circuitry that originally evolved to serve nonmoral functions, such as reward and punishment processing. We will do so by discussing parallels between moral and microeconomic utility theories and by suggesting that classical learning theory, with its principle of error-driven adjustment of reward predictions, can give clues as to how people acquire moral behavior. Neuronally, we will propose that reward regions, such as the striatum and the prefrontal cortex, may contribute to the acquisition and expression of moral behavior. Our proposal implies that the reward system may have beneficial functions not only for individual survival by computing reward value, but also for society and culture by contributing to the computation of the moral value of decisions and actions.

The Is–Ought Gap

When we ponder the morally right course of action, we always face a problem. Philosophers since Hume (1739) have argued that the factual characteristics of the situation will themselves never tell us how we ought to act. Knowing who the hijackers are, what they are going to do, and what the consequences of their actions will be is not enough to deduce what the president ought to do. To use Hume's words, "you cannot derive an ought from an is" (Hume, 1739; for a related argument, see Moore, 1903; for a comparison, see Bruening, 1971; for a dissenting view, see Searle, 1964). In order to find out what the president ought to do, we need not only information about the facts of the current situation, but also some normative claims about what is valuable. We would commit a logical fallacy if we were to deduce moral conclusions just from descriptive facts.

In the airplane example, one possible normative claim might be that we are morally obliged to provide happiness for the largest possible number of people. This would be a utilitarian normative claim (e.g., Mill, 1861). Another possibility would be to presuppose the moral rule that one ought not to kill. This would be a deontological normative claim (e.g., Kant, 1797). Descriptive information about the situation must be combined with such normative claims in order to reach a conclusion about what one ought to do. For example, if one embraces the normative claim that we are morally obliged to save as many people as possible, we could determine what to do by combining this claim with factual information about how different possible actions would affect the number of people saved. Thus, normative claims and descriptive information taken together allow us to make moral decisions.

Moral Theories

We derive normative claims from moral theories. The main classes of moral theories are the utilitarian (consequentialist), deontological, and virtue theories. Utilitarianism states that the moral quality of actions is determined by their consequences (*utilitas* = Latin for [public] good, usefulness). This means that we have the duty to maximize the sum or mean utility of all individuals in the society. Traditionally, utility is interpreted hedonistically as balance of pleasure and pain, but it could also include beauty and truth (Moore, 1903), (informed) preference fulfillment (Brandt, 1979), and any other valuable consequences of actions and rules. For utilitarian theories, something is morally valuable *because* it promotes nonmoral values such as happiness, pleasure, or other kinds of utility.

Deontological theories argue that certain things are morally valuable in themselves. Such moral value does not derive from the fact that it promotes happiness, pleasure, or any other utility (Rawls, 1971). The moral quality of actions arises from the fact that the action is done to protect such moral values. The consequences and outcomes of actions are secondary. Impermissible actions remain so irrespective of how good their consequences are. Some deontological theories claim that all moral values can be traced to one ultimate moral principle. One example of such an ultimate principle is autonomy, conceived as the ability of a person to generate self-binding laws (Kant, 1797). Another example is the ability to act (Gewirth, 1998). Other positions claim that there are several moral values, which have equal standing. Ross (1930), for example, proposes seven basic values, which correspond to seven moral duties: fidelity, reparation, justice, beneficence, gratitude, self-improvement, and noninjury. Kant's (1797) work on morals is the paradigmatic example of a deontological theory. The core rule in Kant's system is the categorical imperative, which requires that we act according to maxims that could become universal laws. Rules often arise from the rights of those affected by the agent's actions (Kamm, 2007). In Kantian terminology, this is expressed by the claim that one ought to treat others as ends in themselves, and never only as means to an end.

Virtue ethics focuses on the motives and character traits of actors. The bearers of moral quality are not actions but life, people, and their capabilities (Nussbaum, 2000; Sen, 1985). Virtues are goods in themselves, and their continued practice contributes to the good life. Examples of virtuous dispositions are wisdom, courage, moderation, justice, kindness, generosity, self-respect, compassion, altruism, forgiveness, and sincerity. Virtues may fall in between more extreme character dispositions. Having too little or too much of the disposition results in vices (e.g., cowardice or rashness for courage; Aristotle, *Nicomachean Ethics*, 2002 ed.). Virtues should be carefully cultivated through moral education by role models. The development of habits helps to make the expression of virtuous dispositions more stable and reliable, but virtuous behavior should nevertheless remain goal directed and purposeful. Applying and reconciling virtues requires practical wisdom, which helps to discern what is important for a good life. Virtues are associated with an appropriate feeling of the right emotions. Aristotle and, more recently, Anscombe (1958), MacIntyre (1985), and Foot (2001) are some of the prime proponents of virtue ethics.

Moral Decisions in Theory and Practice

An implication of the is–ought gap is that whereas moral theories provide standards for how we should act, they do not describe how moral judgments and decisions are achieved in practice. Accordingly, moral theories cannot be falsified by empirical findings that we often do not behave in ways that would be prescribed by the theories. In fact, even if all of our actions and dispositions were to contradict what moral theories require from us, what is morally prescribed would still remain the same. Thus, in further illustration of the is–ought gap, moral theories are about how we ought to behave, not about how we do behave.

Similarly, to answer the question of what we should do does not automatically answer the question of how moral decisions should be achieved (e.g., Bales, 1971). It is, in fact, open to discussion whether moral theories do or do not suggest specific methods for moral decision making. For example, from a Kantian perspective, it is unclear how we should determine whether our maxims could become universal laws. Some deontological theories suggest that we should follow the (juridical) law in order to behave morally (Wikström, 2007). Virtue theories explicitly require deliberate and purposeful application of the positive dispositions. Utilitarian theories would seem to suggest that we should make moral decisions on the basis of a process of calculating utilities. But in fact, several utilitarians actually recommend that we do not calculate moral utility every time before we make a decision (e.g., Mill, 1861; Sidgwick, 1874)—for example, because the required calculations are too difficult and, therefore, error prone or take too much time. Conversely, following our intuitions will often lead to utility-maximizing decisions (Hare, 1981).

The Relevance of Decision-Making Research for Moral Theories

As we have seen above, moral theories are prescriptive, not descriptive. Nevertheless, moral theories cannot prescribe things we cannot possibly live up to. This is often expressed as the *ought-implies-can* principle, which is ascribed to Kant (1788). The idea is that the concept of having a duty already logically entails that the addressee is able to fulfill it. To say that *you ought to do A* already presupposes that *you are able to do A*. The notion of *can* is often interpreted to refer both to ability and to opportunity (Copp, 2008). For example, I can have a duty to save a drowning child only if I am actually capable of saving the child (if I can't swim and would drown myself before reaching the child, I do not have a duty to save it) and if I have the opportunity to do so (if I am in China at the moment, I do not have a duty to save a child drowning in a pond in Manchester). The ought-implies-can principle is also used to defend the claim that there would be no room for moral responsibility if the universe were deterministic (Copp, 1997; Widerker, 1991).

This ought-implies-can principle provides us with a connection with empirical research on decision making. Moral theories presuppose that we can have a duty to perform acts only insofar as we are able to perform them, and to perform the morally right action we must first reach the morally right decision. And that is where the mechanisms of moral decision making enter the picture. Insofar as moral theories suggest certain preferred methods of moral decision making (and we have seen that they sometimes appear to do that—e.g., by advocating intuitive decision making), they must presuppose that these methods enable us, at least in some cases, to make the right decision. Empirical research might tell us whether that is a valid presupposition.

Empirical Findings for Moral Judgment and Decision Making

Research has focused on moral dilemmas that oppose deontological and utilitarian theories. In the dilemma of the runaway trolley (Thomson, 1985), a trolley will kill five workers unless diverted to a sidetrack where it will kill one worker. The majority of people choose to switch the trolley to the sidetrack. The outcome of this choice is in agreement with the utilitarian argument that one ought to maximize overall utility. In an alternative version of the dilemma (the footbridge dilemma), the runaway trolley cannot be switched off the main track. The trolley can be prevented from reaching the five workers only by pushing a fat man from a bridge onto the track, thus killing him. The majority of people refuse to push and kill the fat man in order to save the five workers. One interpretation of these findings (Greene, Nystrom, Engell, Darley, & Cohen, 2004) suggests that the deontological consideration not to use a person purely as a means to an end (Kant, 1797) overrides considerations of utility maximization. This would mean that in real life, people are, at least at first glance, capable of approaching moral issues from both utilitarian and deontological points of view. Moreover, they do not commit themselves to one moral theory but switch between different theories as the circumstances change. People may thus eclectically combine considerations from different moral theories. However, this point is somewhat speculative, since we do not know how participants have actually come to their decision. The mere fact that the outcome of the decision maximizes utility does not make it a utilitarian decision; instead, the decision might be made on deontological grounds or on no grounds at all.

An alternative interpretation of the difference between the two dilemmas suggests that the emotional involvement is higher for the footbridge than for the trolley problem because the act of pushing a person constitutes a more personal and harmful action than does changing a switch (Greene et al., 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; but see Nichols & Mallon, 2006). Given that at least rough dissociations can be made with respect to the functions of different brain regions (e.g., Fuster, 1997), the emotion-based interpretation of the footbridge problem may be endorsed by results from functional neuroimaging and lesion studies. For example, within the prefrontal cortex, ventromedial regions are more strongly involved in the processing of emotions, whereas dorsolateral regions fulfill more cognitive functions (Duncan & Owen, 2000; Fuster, 1997). An emotion-based account of the footbridge problem would therefore predict activation of ventromedial prefrontal regions while participants solve footbridge-like problems. This is indeed what was found (Greene et al., 2001).

Conversely, making decisions in trolley-like problems may draw on more cognitive capacities, in that it involves weighting up costs (e.g., of one man dying) and benefits (e.g., of five men being saved). One may therefore expect a stronger involvement of the dorsolateral prefrontal cortex during trolley-like than during footbridge-like problems, and this is also what was found (Greene et al., 2001).

The dorsolateral prefrontal cortex is also more active with utilitarian judgments in footbridge-like problems when participants accept, rather than reject, personal norm violations if these lead to a substantial increase in happiness (Greene et al., 2004). In summary, these imaging results suggest a dissociation of the ventromedial and dorsolateral prefrontal cortex during the solving of moral problems, with ventromedial regions processing the emotionally salient aspects of footbridge-like problems and dorsolateral regions analyzing costs and benefits.

The functional dissociation of ventromedial and dorsolateral prefrontal activations concurs well with the finding that lesions of the ventromedial prefrontal cortex result in a higher proportion of utilitarian moral judgments in footbridge-like problems, but not in less emotional, trolley-like problems (Koenigs et al., 2007; see also Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005). Without emotional input from the ventromedial prefrontal cortex, the utilitarian reasoning of the dorsolateral prefrontal cortex may dominate behavior in footbridge-like problems. Taken together, these results may suggest that healthy participants weigh up ventromedially processed emotional factors with dorsolaterally processed utility-related factors to determine the morality of a judgment.

The interpretation of the ventromedial lesion results is somewhat complicated by studies of the ultimatum game. In the ultimatum game, a proposer offers a way in which to split a monetary amount, and a responder either accepts or rejects the offer. In the case of acceptance, the amount gets split according to the offer; in the case of rejection, neither player receives anything. As with footbridge-like problems, unfair ultimatum offers (people typically reject offers of about 20% or less of the monetary amount) may elicit a combination of emotional and utility-related processes. Fairness motives, inequality aversion, disgust, and anger may motivate rejection of unfair offers, whereas utility-related factors, such as self-interest and the minimization of personal cost, may motivate their acceptance. Ventromedial patients are more likely to reject unfair offers (Koenigs & Tranel, 2007; for reviews, see Greene, 2007, and Moll & de Oliveira-Souza, 2007; see also Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Thus, with unfair offers in the ultimatum game, ventromedial patients appear to weigh emotional and fairness factors more strongly than utilitarian self-interest, whereas the opposite appears to be the case with footbridge-like problems. Differences in the types of emotion (empathy vs. anger) and utility (welfare of others vs. monetary self-interest) involved in footbridge-like problems and the ultimatum game may explain the apparent discrepancies in the behavior of ventromedial patients.

What Cognitive Processes a Utilitarian Moral Theory May Implicate

From a philosophical perspective, the moral domain is special in that it is concerned with what we ought to do, rather than with what we actually do. Is the moral domain also special from a psychological and neuroscientific perspective, in that there are dedicated mechanisms for moral cognition and decision making, computed in a dedicated class of neurons and structures (for such a view, see Harman, 2000; Hauser, 2007; Mikhail, 2007; for a critique, see Dupoux & Jacob, 2007)? Due to space restrictions, we will focus in the following on utilitarian moral theory. On the basis of (1) parallels of utilitarian moral theory and economic decision theories, (2) findings of neuronal correlates of terms used by economic decision theories, and (3) potential relations with learning theory, we will propose that at least some moral cognition may not be special but, rather, may make use of evolutionary old and proven reward decision mechanisms.

Utilitarians argue that in order to make the morally right decision, we need to determine which action or rule maximizes utility. In order to determine whether a particular action or rule maximizes overall utility, we need to consciously calculate or intuitively estimate utility and, thus, predict and weigh the consequences of an action. We also have to predict how the utility of the people affected will change, which requires that we recognize others as having different preferences and taste. By empathizing and identifying with others, we can determine what it would feel like to be in their situation. We can extrapolate how one specific action or rule implementation, rather than another, will maximize overall utility by stepping into many different shoes and comparing the utilitarian consequences of the action or rule (Harsanyi, 1955). Thus, moral utilitarianism implicates several cognitive and emotional processes and, consequently, provides leads for investigating moral decision making.

In the following section, we will explore the parallels between moral and nonmoral utility-based decisions. As was mentioned in the introduction, there is some disagreement between utilitarians who argue that the right moral decision requires conscious calculation of utility and those who argue that other (more intuitive) methods are just as good or better. This discussion bears remarkable similarities to issues discussed in economic theory. It is possible that we can learn something about moral decision making by looking at parallels with other theories and at findings that support the theories. By extension, we could then investigate whether moral decision making taps into these processes, which have not originally evolved to subserve moral reasoning as such.

Parallels of Moral Utilitarianism With Economic Decision Theory

Utility has a long-standing tradition in economics. Bernoulli (1738) argued that each additional unit of a good, such as money, becomes subjectively less valuable (diminishing marginal utility). An additional $1,000 yields more marginal utility in our poor student times than when we have reached some financial security and wealth. In the moral domain, diminishing marginal utility provides an argument for wealth redistribution and supporting the poor, because it maximizes overall utility (Mill, 1861). Assuming similar utility functions for the poor and rich, the utility loss incurred by the rich from giving to the poor is much smaller than the utility gain of the poor. Modern economic decision theory stresses the notion that utility is content free; thus, the connotation of utility as pleasure and happiness no longer holds. People reveal their preferences in overt choice behavior between two options, and utility theory captures these revealed preferences with a utility function. Preferences differ between individuals, and there are no normative prescriptions of what people ought to prefer.

As long as preferences fulfill some axioms, such as being well ordered, transitive, complete, and independent, a utility function can be defined (Marschak, 1950; Von Neumann & Morgenstern, 1944). We can then obtain the expected utility of a choice option by multiplying the utility (u) of all possible outcome magnitudes (m) of the option with their probabilities (p) and integrating across products: $\text{EUT} = \Sigma[p * u(m)]$. In order to maximize expected utility, we should choose the option with the largest sum. Diminishing marginal utility results in concave (Bernoulli proposed logarithmic; alternatives are other power or quadratic and exponential) utility functions. Concave utility functions imply risk aversion (Bernoulli, 1738). A risk-averse decision maker prefers $500 for sure over the toss of a fair coin for $1,000 or $0, even though the two options have the same expected value. A concave utility function implies that the first $500 yields more marginal utility than does the second $500. Therefore, the sure option will be preferred over the average utility of $1,000 and $0. The more concave the utility function, the more risk averse the agent. Conversely, the more convex the utility function, the more risk seeking the agent, and the more he is willing to accept a risky option with probabilistic high-magnitude outcomes.

The neuronal systems processing economic value partly overlap with those processing moral value. Although a neuronal correlate of a formal economic expected utility signal has not yet been identified, correlates of components such as magnitude and probability are known to be processed by the reward system (for more details on the reward system and its functions, see, e.g., Berridge & Robinson, 2003; Rushworth & Behrens, 2008; Schultz, 2006). Dopamine neurons in the midbrain show phasic activations with short latencies of about 100 msec to unpredicted rewards and to reward-predicting stimuli. These activations scale with the magnitude and probability of predicted reward and combine the two parameters (Tobler, Fiorillo, & Schultz, 2005). It currently remains to be tested whether dopamine neurons code expected value (sum of probability-weighted reward magnitudes), expected utility (sum of probability-weighted utilities of reward magnitudes), prospect (wealth changes, weighted by a distorted probability function), or another combination of magnitude and probability.

Dopamine neurons send widespread projections to many subcortical and cortical regions, particularly the striatum and prefrontal cortex (Gaspar, Stepniewska, & Kaas, 1992). Both striatal and prefrontal neurons code reward magnitude and probability (Cromwell & Schultz, 2003; Kobayashi, Lauwereyns, Koizumi, Sakagami, & Hikosaka, 2002; Wallis & Miller, 2003). Furthermore, neurons in these regions combine reward information with more detailed sensory and motor information and are in a position to compute adaptive behavioral signals with integrated reward value (Lau & Glimcher, 2008; Padoa-Schioppa & Assad, 2006; Sakagami & Watanabe, 2007; Samejima, Ueda, Doya, & Kimura, 2005). In humans, activations scaling with reward magnitude and probability are prominent in the striatum and prefrontal cortex (Abler, Walter, Erk, Kammerer, & Spitzer, 2006; d'Acremont & Bossaerts, 2008; Knutson, Taylor, Kaufman, Peterson, & Glover, 2005; Tobler, O'Doherty, Dolan, & Schultz, 2007; Yacubian et al., 2006). Moreover, risk-related activations in the prefrontal cortex vary with risk attitude, with more lateral prefrontal regions showing stronger risk signals with increasing risk aversion and more medial regions showing them with increasing risk proneness (Tobler et al., 2007). As was explained above, risk attitude relates to the shape of the utility function, with concave functions reflecting risk aversion and convex ones risk proneness. Taken together with the finding that lateral prefrontal regions may weigh up costs and benefits in moral problems and medial prefrontal regions may process the emotional aspects of moral problems (Greene et al., 2004; Knoch et al., 2006; Koenigs et al., 2007), these results suggest that a variety of reward regions contribute to the calculation of an economic utility signal and that some of these regions overlap with those implicated in the processing of moral utility.

The neuronal data on footbridge- and trolley-like problems suggest involvement of at least two interacting systems, one deliberative and one emotional (Greene et al., 2004; Greene et al., 2001). This suggestion is reflected in the neuroeconomic domain, with research proposing one set of systems involved in high-level deliberative processes, such as problem solving, planning, and trading off costs with benefits, and another set of systems engaged by the emotional aspects of the decision, such as impatience, distress of deferring immediate gratification, and fear of loss. Multiple-system proponents in neuroeconomics concur with their moral neuroscience colleagues in locating the more deliberative aspects of decision making within the dorsolateral prefrontal (and parietal) cortex. Conversely, the more affective aspects of decision making are thought to be located within the ventromedial prefrontal cortex (as well as within the insula and subcortical structures, such as the amygdala and striatum; Berns, Laibson, & Loewenstein, 2007; McClure, Ericson, Laibson, Loewenstein, & Cohen, 2007; McClure, Laibson, Loewenstein, & Cohen, 2004; Miller & Cohen, 2001; Sanfey, Loewenstein, McClure, & Cohen, 2006). However, alternative views suggest that the distinction into different neuronal systems and mental processes may be more a matter of gradation than of category. In this view, the brain acts as a single information-processing system (Glimcher, Kable, & Louie, 2007; Kable & Glimcher, 2007; Kalenscher & Pennartz, 2008; Tom, Fox, Trepel, & Poldrack, 2007). The single- versus multiple-system question is the focus of ongoing research (Rustichini, 2008). In the moral domain, avenues for further research may be suggested, in that single-system views presently appear to be more prominent than multiple-system views in theoretical work, whereas the opposite may hold in empirical work.

Expected utility theory can be extended into the social and moral domain with similar formalisms (Harsanyi, 1955). Economic utility functions represent our egoistic preferences with respect to risky economic outcomes; social or moral utility functions represent impersonal preferences over income distributions or social states (Harsanyi, 1953). Thus, social preferences are preferences that we hold irrespective of our own situation and interests. An example would be our view of what a just tax system should look like, irrespective of how much we actually earn. The social utility functions can vary between individuals, just as with personal utility functions. Within an individual, the social and the personal utility functions can be in opposition—for example, with respect to an income distribution that favors the individual. Such a distribution may have low social but high personal utility. To maximize social utility, we should use the same principles and formalism as those proposed by expected utility theory described above (Harsanyi, 1955). By adding up the individual social utilities and maximizing the probability-weighted sum, we determine the income distribution or social state that maximizes the expected social utility of all members of the society.

Personal and social outcomes of decisions could be combined in single utility functions. As a proposer in the ultimatum game, we might offer a relatively high amount to a responder partly because we prefer the responder to get an amount similar to our own. More specifically, a person (i) with inequality aversion might discount their personal gains (x_i) by both self-interested disadvantageous and other-regarding advantageous inequality. In the two-player case with just one other player (j), the utility of player i corresponds to $U_i(x) = x_i - \alpha(x_j - x_i) - \beta_i(x_i - x_j)$ (Fehr & Schmidt, 1999). Disadvantageous inequality results in the utility loss $\alpha(x_j - x_i)$, advantageous inequality in the utility loss $\beta_i(x_i - x_j)$. The two weighting parameters, α and β, could be interpreted as individual propensity for envy (that you get more than I) and compassion (I feel with you because I get more than you), respectively. In most people, α is larger than β. Assuming a utility function that incorporates other-regarding inequality aversion can also explain why we may prefer to cooperate, rather than defect, in the prisoner's dilemma and punish defecting players at a cost to ourselves (Fehr & Camerer, 2007).

We know that reward-processing regions encode not only the value of primary outcomes, such as food and drink, but also that of higher order rewards, such as money, abstract points, cooperation, and fairness (e.g., Delgado, Frank, & Phelps, 2005; Delgado, Labouliere, & Phelps, 2006; Huettel, Stowe, Gordon, Warner, & Platt, 2006; Knutson et al., 2005; Rilling et al., 2002; Singer, Kiebel, Winston, Dolan, & Frith, 2004; Tabibnia & Lieberman, 2007; Yacubian et al., 2006). For example, activation in the ventral striatum increases with the ratio of what one participant receives, as compared with another (Fliessbach et al., 2007). By extension, this social reward comparison output from the ventral striatum could form the basis for a formal inequality aversion signal. Other-regarding motives such as empathy are processed by the anterior insula and the anterior cingulate cortex (Singer, Seymour, et al., 2004; for a review, see Frith, 2007). We suggest that a derived, utilitarian moral value signal could be computed by reward-processing regions in a manner similar to that for a standard reward value signal. It would be interesting to test where and how such a signal is combined with other-regarding processes. One prediction could be that the basic components of the utility signal will be combined in the prefrontal cortex and striatum similarly for economic and moral utility but that moral utility, in addition, may draw on other-regarding structures, such as the insula and cingulate cortex.

Parallels of Moral Utilitarianism With Behavioral Economics

Just like ethical theories, expected utility theory has a normative flavor entailed in the axioms and computations that it requires for optimal behavior. As a consequence, one may want to consider what decision makers can possibly do and what they actually do when discussing normative models in both ethics and economics. Whereas moral theories define optimal behavior as morally good behavior, for expected utility theory it is defined in terms of rationality. Expected utility theory thus prescribes what behavior would be most rational. However, empirically, our preferences sometimes violate the axioms posited by expected utility theory (Allais, 1953; Ellsberg, 1961; Kahneman & Tversky, 1979; Loomes, Starmer, & Sugden, 1991). We may prefer option y over y', but after adding z to both of these options, we may prefer $y' + z$ over $y + z$. Or we may prefer y over y' when the options are described as gains but may prefer y' over y when they are described as losses. Recall the example of airplane hijackers from the beginning. Imagine that they are expected to kill 600 people. We have to choose between two options to combat them. When framed in terms of gains, either 200 people will be saved for certain (safe, risk-free option), or there is a 1/3 chance that 600 people will be saved and a 2/3 chance that no one will be saved (risk-seeking option). When framed in terms of losses, 400 people will die for certain (risk-free option), or there is a 1/3 chance that no one will die and a 2/3 chance that 600 people will die (risk-seeking option). As a variety of different studies have shown (reviewed in Kuhberger, 1998), participants tend to choose the safe option (200 saved for certain) when the problem is framed in terms of gains. Conversely, they choose the risk-seeking option (1/3 chance that no one will die and a 2/3 chance that 600 will die) when the problem is framed in terms of losses. However, such behavior can not be accommodated by expected utility theory, because the two versions of the problem differ only in how the outcomes are described; the actual numbers of dead or alive people are the same.

Functional neuroimaging suggests that amygdala activity is susceptible to economic versions of the framing effect, whereas stronger activation of the medial and orbital prefrontal cortex correlates with stronger resistance to framing (De Martino, Kumaran, Seymour, & Dolan, 2006). On the basis of these findings, an obvious prediction would be that patients with lesions in medial and orbital prefrontal regions would be more susceptible, and amygdala patients less susceptible, to framing in both the moral and the economic domains. Interestingly, people with unconditional moral values as proposed by deontological theories are less susceptible to framing effects in the moral domain (Tanner & Medin, 2004). These people appear to pay more attention to the question of whether their actions are compatible with their moral principles than to the consequences of their actions (to the amount of utility it generates). However, it should be noted that most people are willing to soften even their most protected unconditional values if the consequences of rigidly sticking to them would be too outrageous (Baron & Ritov, in press). In any case, amygdala activity appears to contribute to flexibly adjusting behavior with economic frames, and it might be interesting to investigate whether this holds also with moral framing.

The framing effect contradicts standard economic theory. In response to such contradictions, which usually boil down to axiom violations, one can either relax some of the axioms (e.g., Fishburn, 1982; Machina, 1982) or pursue a nonaxiomatic approach (Kahneman & Tversky, 1979; cf. Kalenscher & Pennartz, 2008). Nonaxiomatic approaches describe actual, rather than deduce optimal, choice behavior. For example, Kahneman and Tversky empirically investigated human choice behavior and found several consistent and systematic deviations from the normative ideal dictated by expected utility theory. These observations are bundled in prospect theory, which, contrary to expected utility theory, submits that we overweigh small and underweigh large probabilities, are usually risk-seeking and not risk averse for losses, and are more sensitive to losses than to gains. However, the computational core of prospect theory and expected utility theory is very similar: According to both theories, value is computed as the sum of utilities of all possible outcomes, weighted by distorted or linear probability. People assign these values to each available alternative and choose whichever alternative yields the highest value. Thus, some basic mechanism of combining probability and magnitude and its neuronal correlates may also be relevant for non- or less axiomatic theories, in both the economic and the moral domains.

Economic and Moral Heuristics

Expected utility theory and moral utilitarianism have often been criticized as requiring excessive processing and calculation (multiplying probabilities and magnitudes and summing the products). Current theories of decision making therefore often propose that humans use simple and straightforward decision rules (heuristics) to make economic or moral decisions. These heuristics are usually effective and put little demand on cognitive processing but can occasionally lead to systematic biases, misjudgments, and deviations from the normative standard (Tversky & Kahneman, 1974). On the other hand, heuristics may have high descriptive validity while performing equally well, or better, in complex environments, as compared with intricate normative models. Examples in the economic domain include the *satisficing* (Simon, 1955, 1956) and the *take-the-best* algorithms (Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002). We often *satisfy* a simple choice criterion that is part of a sequence of hierarchically organized criteria and *suffice*—that is, stop the decision process—once a criterion has been met. For example, we may search only for the option yielding the fastest outcome, while overlooking other, slower options that would potentially yield higher outcomes. Thereby, we satisfy the *minimize waiting time* rule and suffice once the option with the shortest waiting time has been found. Satisficing has also been introduced in utilitarian moral theory (Slote, 1984). The notion deals with the objection that utility maximization cannot be a norm for daily moral behavior. Instead, we should satisfice: aim to increase utility only to a certain level. So, more generally, we may use similar heuristics for making moral and economic decisions.

It is currently a point of contention whether heuristics are near-optimal tools that evolved for optimal decisions, given cognitive and environmental constraints, or empirical deviations from normative economic and moral theories (Gigerenzer & Goldstein, 1996; Sunstein, 2005). Like economic heuristics, moral heuristics may be sufficient most of the time but can fail or result in inconsistent behavior with difficult decisions, such as the trolley and the footbridge problems. Examples of specifically moral heuristics are *do not knowingly cause a human death*, *people should not be permitted to engage in moral wrongdoing for a fee*, *punish and do not reward betrayals of trust*, *penalties should be proportional to the outrageousness of the act*, *do not play God or tamper with nature*, and *harmful acts are worse than harmful omissions*. The rule-of-thumb-like nature of these heuristics might give them a deontological flavor, but it is possible to remain within a strictly utilitarian framework: Rule-utilitarianism has argued that we ought to follow utility-maximizing rules (see, e.g., Brandt, 1967; Hooker, 2000). In such a view, the moral quality of actions depends on whether they accord with a rule that would maximize the sum or mean utility if everybody were to follow it.

Although people often make decisions consistent with utilitarianism, they sometimes maximize utility only of members of a certain group (e.g., fellow citizens, coworkers, etc.), but not of members of another group (e.g., citizens of another country, workers from another factory, etc.). Such *parochialism* (Baron, in press) refers to the tendency of a decision maker to accept personal costs to maximize the benefit of other in-group members, while neglecting or even disproportionally increasing the negative effects on outsiders (for experimental evidence, see Bornstein & Ben-Yossef, 1994). As a consequence, the net overall utility (when everyone is taken into account) is reduced. Parochialism seems to be a moral heuristic in itself (*protect the interests of my fellow in-group members*), which would suggest that we make some utilitarian decisions only with respect to the groups we belong to.

If it is unclear why, when, and whether ethical intuitions maximize utility, one may ask why we all seem to use them and why they have developed in the first place. Moral heuristics, including group loyalty and parochialism, may have provided an evolutionary advantage to our ancestors and could have evolved from simple social and emotional responses (Greene, 2003; Greene et al., 2004). For example, it is certainly beneficial for the prosperity and evolutionary stability of a society if many of its members follow certain moral rules of thumb, such as *don't kill* or *respect the personal property and integrity of in-group members*. Such social instincts presumably evolved in situations that primarily involved direct personal contact (small, closed societies), but not in environments in which spatially and temporally distant strangers helped each other. This can explain why we usually feel more obliged to help individuals that are presented to us in flesh and blood than to help remote, abstract, and temporally and spatially distant individuals. This evolved social rule obliges us not to kill the fat man with our own hands by throwing him onto the train tracks in the footbridge dilemma. But because this heuristic does not apply if we are not face to face with the fat man, it does not sufficiently discriminate between the choice alternatives in the scenario in which we manipulate the track switch from a remote control room. Thus, we rely on a second rule of thumb: Try to save as many lives as possible by minimizing the number of deaths. A satisficing algorithm can thus provide an alternative explanation for the ethical eclecticism we described earlier: Perhaps people are not sometimes utilitarians and sometimes deontologists but use rules of thumb to make decisions that are essentially about utility. This does not mean that people never make use of deontological moral principles; it means only that the use of rules does not necessarily imply such deontological principles. As was outlined above, the hallmark of deontology is its reference to values that are themselves moral in nature. Whether or not a rule is a *deontological* rule thus depends on the nature of the values the rule refers to. Someone who employs the *don't kill* rule can justify that rule on a utilitarian basis (such a rule brings about the most happiness), but also on a deontological basis (life has moral value and thus must be protected). In the first case, the rule is justified by referring to values that are themselves not moral (such as happiness); in the second case, the underlying values are themselves moral.

Can We Learn to Decide More Morally?

From both a personal and a societal point of view, we might be interested in how we acquire morality and whether it is possible to make better moral decisions. Building on Piaget (1932), Kohlberg proposed that moral development consists of the stagewise acquisition of a deontological moral theory (for a review, see Colby & Kohlberg, 1987). Kohlberg ascribed moral theories to children on the basis of the justifications they gave for their moral decisions. He identified six stages of moral development. At the first stage, the child is oriented to obedience and punishment. The concern is with what authorities and the law permit and punish. At the second stage, the child is concerned with self-interest and fair exchange. The third stage is reached at the beginning of the teenage years and is concerned with the motives of actions, interpersonal relations, and emotions such as empathy and trust. The fourth stage is about duty to keep social order, respect authorities, and obey the law because of its role for society. At the fifth stage, people reason about social contracts, the democratic process, and human rights. The sixth stage concerns the universal principles that achieve justice, which involves impartially taking the perspective of all parties and respecting them equally. According to the theory, we move from one stage to the next by thinking about moral problems, perhaps after our viewpoints have been challenged by others.

Views from learning theory. Kohlberg's theory has been criticized for putting deontological over other moral theories (e.g., Campbell & Christopher, 1996). Moreover, it is concerned with abstract thinking about moral situations, which might differ from action. Are there alternative mechanisms that could explain the acquisition of moral behavior and real-life decision making? One possibility may come from learning theory. Cultural differences in ethical values suggest that moral reasoning may indeed be learned. Learning theory aims to explain how rewards and punishments modify behavior, and how learning agents form associations between such reinforcing stimuli and actions or conditioned stimuli. Rewards and punishments can be survival related, as is the case with drink, food, sex, and pain, or of a derived, higher order nature, as is the case with monetary gains and losses, social recognition, and contempt. It is conceivable that we learn to decide morally by associating actions and stimuli with social and nonsocial rewards and punishments. Through higher order conditioning, we may have learned that sometimes it takes a very long time until a reward occurs. For example, we may learn to punish individuals who violate social norms (de Quervain et al., 2004; Fehr & Gächter, 2002) or to trust others (King-Casas et al., 2005) because this will lead to more reward for us in the long run. On the other hand, deciding morally and following a moral principle might be rewarding in itself. As has been mentioned, to increase the utility of others may increase our own utility (Fehr & Schmidt, 1999). In summary, not only learning about reinforcement value, but also learning about moral value may follow principles from learning theory. In the next section, we therefore will briefly introduce a prominent learning theory and explore its applications to the moral domain.

Modern theories of conditioning. Modern learning theories (for a review, see Dickinson, 1980) describe association formation as a function of the degree to which participants process the stimuli and their reinforcing outcomes. According to Rescorla and Wagner (1972; see also Sutton & Barto, 1981, 1990), learning of an association depends on the extent to which the outcomes are surprising or elicit a prediction error. Learning is captured as changes in associative strength between the stimuli and the outcomes: $\Delta V = \alpha\beta(\lambda - \Sigma V)$, where ΔV denotes the change in associative strength of a stimulus in a trial; α and β denote the salience or intensity of the stimulus and the outcome, respectively; λ corresponds to the asymptotic processing of the outcome when it is completely unpredicted; and ΣV denotes the sum of the associative strengths of all the stimuli present for the outcome. $\lambda - \Sigma V$ corresponds to the error in prediction of the outcome and can assume positive or negative values in a bidirectional fashion.

Actions and stimuli do not always lead to the same reinforcing outcomes. Outcome distributions can be characterized by their mean (first moment), by their variance (second moment), and by higher moments. Traditionally, prediction errors are computed for the mean reward (first moment). However, the mechanism can, in principle, also be used to learn about higher moments. The activation profiles of single dopamine neurons and of the striatal BOLD response correspond to a prediction error signal for mean reward expectation, possibly normalized by the standard deviation (square root of variance) of the prediction (McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Tobler et al., 2005; Waelti, Dickinson, & Schultz, 2001). Formal variance prediction signals may occur in structures such as the insula (Knutson & Bossaerts, 2007), the cingulate (Brown & Braver, 2007), the posterior parietal cortex (Huettel et al., 2006), and the prefrontal cortex (Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005; Tobler et al., 2007). Errors in the prediction of mean and variance of behavior could possibly be used to determine deviations of actual behavior from behavior as it would be expected if it were to follow a normative moral standard (Montague & Lohrenz, 2007). Accordingly, such prediction errors may motivate punishing actions, which, in turn, may be reinforced by observing increased norm conformity of the punished. Errors in the prediction of distress and empathy experienced by others or oneself may elicit moral learning and reduce actions with consequences harmful to others (Blair, 2007). Thus, the same mechanisms as those used for learning about reward and punishment may serve for learning about morally relevant emotions. In return, self-evaluating moral emotions, such as shame, guilt, embarrassment, and moral pride, themselves generate punishment and reinforcement and, thus, an occasion to learn. After learning, we can predict what emotions we are likely to experience for each alternative course of action and can decide accordingly.

When a neutral stimulus predicts a previously conditioned stimulus, the neutral stimulus will be conditioned as well (Rescorla, 1980). Importantly, such second-order and even higher order associations can occur without ever directly pairing the to-be-learned stimuli with reinforcement. Higher order conditioning can also be explained with prediction errors, as is suggested by a real-time extension of the Rescorla–Wagner model (Sutton & Barto, 1990). Activation in the ventral striatum and the anterior insula reflects second-order conditioning of pain (Seymour et al., 2004). It is possible that higher order conditioning plays a role in moral decision making because it allows us to bridge extended periods of time without reward. In seemingly other-regarding behavior such as trust and inequity aversion, motives of self-interest may play a larger role than has previously been assumed, as long as these behaviors eventually lead to more reward.

We learn not only through our own experience, but also from others. For example, by observing others undergoing pain conditioning and following others' instructions, we are in a position to learn about pain-predicting stimuli without having to experience the aversive outcome ourselves (Berber, 1962; Mineka & Cook, 1993; Phelps et al., 2001). The behavior and facial expressions of others and the meaning of the instructions can elicit prediction errors, and standard learning mechanisms may apply (Lanzetta & Orr, 1980). Both observed and instructed pain conditioning elicit amygdala activation, similar to self-experienced pain conditioning (reviewed in Olsson & Phelps, 2007). In addition, regions coding empathy, such as the anterior insula and the anterior cingulate cortex, and language regions contribute to observational and instructed learning, respectively.

In the moral domain, both observational and laboratory evidence suggests that significant others pass moral rules on to children, or at least provide instances for children to extract moral rules. The children subsequently incorporate the rules into their moral decisions (reviewed in Darley & Shultz, 1990). Observational learning without having to directly experience punishment can occur in the classroom, by watching television, by taking part in pretend play episodes, and by observing parents interact with siblings. Moreover, knowledge about the moral status of others modulates the occurrence of prediction errors in the trust game (Delgado et al., 2005). In the trust game, a proposer decides how much of an initial endowment they want to pass on to a responder. The amount passed to the responder will be multiplied (e.g., tripled) by the experimenter before the responder decides how much to pass back to the proposer. In the experiment of Delgado et al. (2005), participants assumed the role of proposer and played with three different fictional responders: one portrayed as morally praiseworthy and good, one as neutral, and one as bad. Irrespective of moral status, the responders passed back nothing or half of the tripled amount, so that the proposers earned the same amount of money with all three responders. The proposers experienced a negative prediction error when the responders passed back nothing or a positive prediction error when the responders passed back half of the money. One would therefore expect prediction-error-related activation in the striatum, as described previously (McClure et al., 2003; O'Doherty et al., 2003). Activation in the dorsal striatum indeed discriminated strongly between positive and negative prediction errors with the neutral responder. However, the discrimination was much weaker or even absent with the bad and the good responders. Throughout the experiment, the proposers continued to invest more money when they played with the good rather than with the bad responder. These data suggest that moral perceptions can alter behavioral and neuronal reward-learning mechanisms. Thereby, the findings further reinforce the main notion of the present article, that moral decision making may rely on, and interact with, reward- (and punishment-)processing circuitry.

Conclusions

The reward system and its neuronal basis have commonly been portrayed as amenable to hijacking by such disastrous substances and practices as drugs of abuse, pathological gambling, and irrational risk taking (for a review, see, e.g., Kauer & Malenka, 2007). Here, we propose that at least in the domain of utilitarian moral value, the reward system could be put to more beneficial use, in that it may help compute moral value. As exemplified in dopamine neurons, the striatum, and the lateral prefrontal cortex, the reward system may combine the probability and magnitude of valuable consequences associated with action options. Formal learning theory suggests mechanisms by which utilitarian moral value may come to guide behavior.

But what about deontological moral theories? These theories presuppose that we value certain moral principles or acts or things *in themselves*, regardless of the amount of utility they produce. How should we understand these kinds of value judgments? Insofar as they take other things than consequences into account, they cannot be related to reward, because reward (at least at face value) is a consequence of actions or preceding stimuli. So the reward system is unlikely to underlie such judgments. It seems hard to conceive any other system that could, without introducing a separate *moral motivation system*. If it turned out that it is impossible for human beings to value things regardless of utility-related characteristics (if human beings can value things only by relating them to reward), this might present deontological moral theories with a problem. After all, the ought-implies-can principle states that human beings should, at least in principle, have the capacity to make the morally right decision. If human beings were unable to ascribe purely deontological value to actions without any regard for utility, this would be a strong objection against deontological theories.

AUTHOR NOTE

This work was supported by the Wellcome Trust, the Swiss National Science Foundation, the Roche Research Foundation, and Senter Novem. The collaboration between the authors was made possible by a grant from the Volkswagen Foundation, and the article has been written as part of the European Platform for Life Sciences, Mind Sciences, and the Humanities initiative of the Volkswagen Foundation. Correspondence concerning this article should be addressed to P. N. Tobler, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY, England (e-mail: pnt21@cam.ac.uk).

REFERENCES

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, **31**, 790-795.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica*, **21**, 503-546.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy*, **33**, 1-19.

Aristotle (2002). *Nicomachean ethics* (S. Broadie, Ed., & C. Rowe, Trans.). Oxford: Oxford University Press.

Bales, R. E. (1971). Act-utilitarianism: Account of right-making characteristics or decision-making procedures? *American Philosophical Quarterly*, **8**, 257-265.

Baron, J. (in press). Parochialism as a result of cognitive biases. In R. Goodman, D. Jinks, & A. K. Woods (Eds.), *Understanding social action, promoting human rights*. New York: Oxford University Press.

Baron, J., & Ritov, I. (in press). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *The psychology of learning and motivation: Moral judgment and decision making* (Vol. 50). San Diego: Academic Press.

Berber, S. M. (1962). Conditioning through vicarious instigation. *Psychological Review*, **69**, 450-466.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, **5**, 175-192.

Berns, G. S., Laibson, D., & Loewenstein, G. (2007). Intertemporal choice—toward an integrative framework. *Trends in Cognitive Sciences*, **11**, 482-488.

Berridge, K. C., & Robinson, T. E. (2003). Parsing reward. *Trends in Neurosciences*, **26**, 507-513.

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

Bornstein, G., & Ben-Yossef, M. (1994). Cooperation in intergroup and single-group social dilemmas. *Journal of Experimental Social Psychology*, **30**, 52-67.

Brandt, R. B. (1967). Some merits of one form of rule-utilitarianism. *University of Colorado Studies in Philosophy*, **3**, 39-65.

Brandt, R. B. (1979). *A theory of the good and the right*. New York: Oxford University Press.

Brown, J. W., & Braver, T. S. (2007). Risk prediction and aversion by anterior cingulate cortex. *Cognitive, Affective, & Behavioral Neuroscience*, **7**, 266-277.

Bruening, W. H. (1971). Moore and "is–ought." *Ethics*, **81**, 143-149.

Campbell, R. L., & Christopher, J. C. (1996). Moral development theory: A critique of its Kantian presuppositions. *Developmental Review*, **16**, 1-47.

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment: Vol. 1. Theoretical foundations and research validation.* New York: Cambridge University Press.

Copp, D. (1997). Defending the principle of alternative possibilities: Blameworthiness and moral responsibility. *Nous*, **31**, 441-456.

Copp, D. (2008). "Ought" implies "can" and the derivation of the principle of alternative possibilities. *Analysis*, **68**, 67-75.

Cromwell, H. C., & Schultz, W. (2003). Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *Journal of Neurophysiology*, **89**, 2823-2838.

d'Acremont, M., & Bossaerts, P. (2008). Neurobiological studies of risk assessment: A comparison of expected utility and mean–variance approaches. *Cognitive, Affective, & Behavioral Neuroscience*, **8**, 363-374.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, **41**, 525-556.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, **8**, 1611-1618.

Delgado, M. R., Labouliere, C. D., & Phelps, E. A. (2006). Fear of losing money? Aversive conditioning with secondary reinforcers. *Social Cognitive & Affective Neuroscience*, **1**, 250-259.

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision making in the human brain. *Science*, **313**, 684-687.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science*, **305**, 1254-1258.

Dickinson, A. (1980). *Contemporary animal learning theory*. Cambridge: Cambridge University Press.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, **23**, 475-483.

Dupoux, E., & Jacob, P. (2007). Universal moral grammar: A critical appraisal. *Trends in Cognitive Sciences*, **11**, 373-378.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, **75**, 643-669.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences*, **11**, 419-427.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, **415**, 137-140.

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

Fishburn, P. C. (1982). Nontransitive measurable utility. *Journal of Mathematical Psychology*, **26**, 31-67.

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C. E., & Falk, A. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science*, **318**, 1305-1308.

Foot, P. (2001). *Natural goodness*. Oxford: Oxford University Press, Clarendon Press.

Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society B*, **362**, 671-678.

Fuster, J. M. (1997). *The prefrontal cortex. Anatomy, physiology, and neuropsychology of the frontal lobe.* Philadelphia: Lippincott-Raven.

Gaspar, P., Stepniewska, I., & Kaas, J. H. (1992). Topography and collateralization of the dopaminergic projections to motor and lateral prefrontal cortex in owl monkeys. *Journal of Comparative Neurology*, **325**, 1-21.

Gewirth, A. (1998). *Self-fulfillment*. Princeton, NJ: Princeton University Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, **103**, 650-669.

Glimcher, P. W., Kable, J., & Louie, K. (2007). Neuroeconomic studies of impulsivity: Now or just as soon as possible? *American Economic Review*, **97**, 142-147.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, **109**, 75-90.

Greene, J. D. (2003). From neural "is" to moral "ought": What are the moral implications of neuroscientific moral psychology? *Nature Reviews Neuroscience*, **4**, 846-849.

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

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, **44**, 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, **293**, 2105-2108.

Hare, R. M. (1981). *Moral thinking*. Oxford: Oxford University Press, Clarendon Press.

Harman, G. (2000). *Explaining value and other essays in moral philosophy*. Oxford: Oxford University Press.

Harsanyi, J. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, **61**, 434-435.

Harsanyi, J. (1955). Cardinal welfare, individualistic ethics, and the interpersonal comparison of utility. *Journal of Political Economy*, **63**, 309-321.

Hauser, M. D. (2007). What's fair? The unconscious calculus of our moral faculty. *Novartis Foundation Symposium*, **278**, 41-50.

Hooker, B. (2000). *Ideal code, real world: A rule-consequentialist theory of morality*. Oxford: Oxford University Press.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision making. *Science*, **310**, 1680-1683.

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., & Platt, M. L. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron*, **49**, 765-775.

Hume, D. (1739). *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. London: Noon.

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

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, 263-291.

Kalenscher, T., & Pennartz, C. M. A. (2008). Is a bird in the hand worth two in the future? The neuroeconomics of intertemporal decision making. *Progress in Neurobiology*, **84**, 284-315.

Kamm, F. M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harms.* Oxford: Oxford University Press.

Kant, I. (1788). *Critik der practischen Vernunft.* Riga: Johann Friedrich Hartknoch.

Kant, I. (1797). *Die Metaphysik der Sitten in 2 Theilen*. Königsberg: Nicolovius.

Kauer, J. A., & Malenka, R. C. (2007). Synaptic plasticity and addiction. *Nature Reviews Neuroscience*, **8**, 844-858.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, **308**, 78-83.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, **314**, 829-832.

Knutson, B., & Bossaerts, P. (2007). Neural antecedents of financial decisions. *Journal of Neuroscience*, **27**, 8174-8177.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience*, **25**, 4806-4812.

Kobayashi, S., Lauwereyns, J., Koizumi, M., Sakagami, M., & Hikosaka, O. (2002). Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex. *Journal of Neurophysiology*, **87**, 1488-1498.

Koenigs, M., & Tranel, D. (2007). Irrational economic decision making after ventromedial prefrontal damage: Evidence from the ultimatum game. *Journal of Neuroscience*, **27**, 951-956.

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

Kuhberger, A. (1998). The influence of framing on risky decision: A meta-analysis. *Organizational Behavior & Human Decision Processes*, **75**, 23-55.

Lanzetta, J. T., & Orr, S. P. (1980). Influence of facial expressions on the classical conditioning of fear. *Journal of Personality & Social Psychology*, **39**, 1081-1087.

Lau, B., & Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, **58**, 451-463.

Loomes, G., Starmer, C., & Sugden, R. (1991). Observing violations of transitivity by experimental methods. *Econometrica*, **59**, 425-439.

Machina, M. J. (1982). "Expected utility" analysis without the independence axiom. *Econometrica*, **50**, 277-323.

MacIntyre, A. (1985). *After virtue*. London: Duckworth.

Marschak, J. (1950). Rational behavior, uncertain prospects, and measurable utility. *Econometrica*, **18**, 111-141.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, **38**, 339-346.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *Journal of Neuroscience*, **27**, 5796-5804.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, **306**, 503-507.

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

Mill, J. S. (1861). Utilitarianism. *Fraser's Magazine for Town & Country*, **64**, 391-406.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Reviews in Neuroscience*, **24**, 167-202.

Mineka, S., & Cook, M. (1993). Mechanisms involved in the observational conditioning of fear. *Journal of Experimental Psychology: General*, **122**, 23-38.

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, **11**, 319-321.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). Opinion: The neural basis of human moral cognition. *Nature Reviews Neuroscience*, **6**, 799-809.

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: Norm violations and their enforcement. *Neuron*, **56**, 14-18.

Moore, G. E. (1903). *Principia ethica*. Cambridge: Cambridge University Press.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, **100**, 530-542.

Nussbaum, M. C. (2000). *Women and human development: The capabilities approach*. New York: Cambridge University Press.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, **38**, 329-337.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience*, **10**, 1095-1102.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, **441**, 223-226.

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., & Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nature Neuroscience*, **4**, 437-441.

Piaget, J. (1932). *The moral judgment of the child*. London: Routledge & Kegan Paul.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Rescorla, R. A. (1980). *Pavlovian second-order conditioning: Studies in associative learning*. Hillsdale, NJ: Erlbaum.

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

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., & Kilts, C. (2002). A neural basis for social cooperation. *Neuron*, **35**, 395-405.

Ross, W. D. (1930). *The right and the good*. Oxford: Oxford University Press, Clarendon Press.

Rushworth, M. F., & Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, **11**, 389-397.

Rustichini, A. (2008). Dual or unitary system? Two alternative models of decision making. *Cognitive, Affective, & Behavioral Neuroscience*, **8**, 355-362.

Sakagami, M., & Watanabe, M. (2007). Integration of cognitive and motivational information in the primate lateral prefrontal cortex. In B. W. Balleine, K. Doya, J. O'Doherty, & M. Sakagami (Eds.), *Reward and decision making in corticobasal ganglia networks* (Annals of the New York Academy of Sciences, Vol. 1104, pp. 89-107). New York: New York Academy of Sciences.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, **310**, 1337-1340.

Sanfey, A. G., Loewenstein, G., McClure, S. M., & Cohen, J. D. (2006). Neuroeconomics: Cross-currents in research on decision making. *Trends in Cognitive Sciences*, **10**, 108-116.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision making in the ultimatum game. *Science*, **300**, 1755-1758.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, **57**, 87-115.

Searle, J. (1964). How to derive "ought" from "is." *Philosophical Review*, **73**, 43-58.

Sen, A. (1985). Well-being, agency and freedom. *Journal of Philosophy*, **82**, 169-221.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., et al. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, **429**, 664-667.

Sidgwick, H. (1874). *The methods of ethics*. London: Macmillan.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, **69**, 99-118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, **63**, 129-138.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron*, **41**, 653-662.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, **303**, 1157-1162.

Slote, M. (1984). Satisficing consequentialism. *Proceedings of the Aristotelian Society*, **58**, 139-163.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral & Brain Sciences*, **28**, 531-542.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, **88**, 135-170.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497-537). Cambridge, MA: MIT Press.

Tabibnia, G., & Lieberman, M. D. (2007). Fairness and cooperation are rewarding: Evidence from social cognitive neuroscience. In C. Senior & M. J. R. Butler (Eds.), *The social cognitive neuroscience of organizations* (Annals of the New York Academy of Sciences, Vol. 1118, pp. 90-101). New York: New York Academy of Sciences.

Tanner, C., & Medin, D. L. (2004). Protected values: No omission bias and no framing effects. *Psychonomic Bulletin & Review*, **11**, 185-191.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, **94**, 1395-1415.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, **307**, 1642-1645.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of Neurophysiology*, **97**, 1621-1632.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision making under risk. *Science*, **315**, 515-518.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, **412**, 43-48.

Wallis, J. D., & Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, **18**, 2069-2081.

Widerker, D. (1991). Frankfurt on "ought implies can" and alternative possibilities. *Analysis*, **51**, 222-224.

Wikström, P.-O. (2007). In search of causes and explanations of crime. In R. King & E. Wincup (Eds.), *Doing research on crime and justice* (pp. 117-139). Oxford: Oxford University Press.

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D. F., & Büchel, C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *Journal of Neuroscience*, **26**, 9530-9537.

(Manuscript received March 8, 2008;
revision accepted for publication July 7, 2008.)