Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

From moral concern to moral constraint

Fiery Cushman

Current research into the neural basis of moral decision-making endorses a common theme: The mechanisms we use to make value-guided decisions concerning each other are remarkably similar to those we use to make value-guided decisions for ourselves. In other words, moral decisions are just another kind of ordinary decision. Yet, there is something unsettling about this conclusion: We often feel as if morality places an absolute constraint on our behavior, in a way unlike ordinary personal concerns. What is the neural and psychological basis of this feeling of moral constraint? Several models are considered and outstanding questions highlighted.

Addresses
Department of Psychology, Harvard University, United States

Corresponding author: Cushman, Fiery (cushman@fas.harvard.edu)

Current Opinion in Behavioral Sciences 2015, **3**:58–62

This review comes from a themed issue on **Social behavior**

Edited by **Molly J Crockett** and **Amy Cuddy**

For a complete overview see the Issue and the Editorial

Available online 17th January 2015

doi:10.1016/j.cobeha.2015.01.006

Moral decisions are hard to make and fun to study. Suppose a woman notices $20 laying by the shoes of a stranger at the front of a checkout line. Her eyes linger on the orphaned bill. Will she point out the money to the customer who may have dropped it, or wait a moment until it can be discreetly pocketed? Watching this moment of uncertainty imparts a vicarious thrill because, to varying degrees, her competing motives are shared by us all.

Psychology and neuroscience have much to say about her motive to keep the money. In fact, the integration of computational, neurobiological and psychological models to explain value-guided learning and choice stands out as one of the foremost accomplishments of contemporary behavioral research [1]. Remarkable efforts have also been made to understand the competing moral value: her desire to return the money. The basic upshot is that the value we place on moral behavior is much like the value we place on $20 [2•]. It is encoded in similar neural structures [3], and integrated into decisions using basically similar processes [4,5••].

Yet this conclusion sits uncomfortably alongside philosophical theorizing, psychological evidence, and ordinary experience. Moral values appear to differ in some very fundamental ways from the prudential value of money, food, companionship, and so forth. Our moral values feel more important, universal, and inviolable [6] — we have the sense that you just *have* to return to the $20, but not that you just *have* to keep it for yourself. Are these differences real? Are they reflected in the neural mechanisms that support moral decision-making? And if so, then how?

What is morality?

Attempts to define morality typically focus on two candidate features. The first is concern for others' welfare, which is emphasized in utilitarian or consequentialist philosophical theories. The second key feature is the concept of an absolute constraint, rule or law. This approach finds its philosophical apogee in the work of Kant.

Following the lead of some philosophers, we could seek to refine a single and exact definition of the moral domain. This is a promising avenue if we wish to spend centuries gridlocked in intractable and often arcane debate. Recently, however, psychologists have charted a different course by arguing that moral cognition comprises multiple distinct but interrelated mechanisms [7,8]. On the one hand, we can and do make flexible tradeoffs between our concern for others and for ourselves; In fact, tradeoffs between ourselves and others exhibit such consistency that the 'welfare tradeoff ratio' is championed by some as the computational core of the moral domain [10]. On the other hand, many acts that do not involve obvious welfare considerations at all are nevertheless widely considered immoral (for instance, consensual sibling incest) [9].

Thus, research into the neuroscience of morality faces at least two big questions. First, what mechanisms acquire and encode moral concern: the value of others' welfare, ultimately allowing us to make decisions that flexibly trade off between interests when they collide? Second, what mechanisms acquire and encode the sense of moral constraint: the representation and value of a moral rule, or law? We have an impressive grip on the first issue, but are startling empty-handed on the second.

Moral concern

There are two principle literatures on the neuroscience of other-oriented concern. One interrogates the neural substrates of the perception of pain or reward in others — that is, the basis of empathy. The second interrogates the

neural substrates of decision-making on behalf of others. Both of these literatures converge on a common conclusion: The mechanisms we use to encode value and make decisions for ourselves are largely overlapping with those we use for others.

The affective experience of pain or otherwise unpleasant experience activates a characteristic network of brain regions including anterior cingulate cortex and anterior insula, along with brainstem and regions of the cerebellum. Numerous studies show a similar network of activation (although not perfectly identical) when people observe pain in others [11••]. Similarly, much evidence suggests that people experience vicarious reward when they see others experience positive outcomes. Regions throughout the dopamine reward network, widely observed to respond to the experience of surprising personal rewards, are also activated when individuals see others experience rewarding outcomes, especially for socially close targets [12••,13]. Finally, researchers have investigated the neural mechanisms involved in making choices for others [14–17], including in situations where this generosity carries a personal cost [4,18,19]. Here, again, the typical finding is that people use similar neural mechanisms when making value-guided decisions for others as they do when making value-guided decisions for themselves [16,20,21].

Moral constraint

In contrast to the well-developed literature on welfare concerns, we know little about how the brain represents moral rules as absolute constraints on behavior. Current research does, however, offer two promising approaches. One possibility is that our sense of inviolable moral rules comes from a unique kind of value representation principally designed to guide our own decision-making. Another possibility is that moral rules are grounded in psychological mechanisms principally designed to judge the actions of others.

Model-free moral values

A dominant theme of research in the last decade is that our sense of moral constraint derives from a unique kind of value representation — that strong rules are grounded in strong feelings. According to one early and influential proposal, the dual process model, controlled cognitive processes are responsible for utilitarian-like assignment of value to welfare while affective processes are responsible for the sense of inviolable constraint on 'up-close and personal' harms [7]. Although certain elements of this model are contested on conceptual [22] and empirical [23] grounds, a wealth of data favors the broad distinction between psychological mechanisms that deliver competing responses in dilemmas pitting general welfare against direct harm [24].

Two recent proposals attempt to translate this insight into the language of contemporary computational cognitive models of decision-making [25,26]. They leverage one of the oldest distinctions in the history of psychology, between goal-directed and habitual action [27]. Goal-directed actions require a working model of the world. You pick a desirable outcome, and then form a plan to bring it about. Thus, they correspond to the class of model-based reinforcement learning algorithms. In contrast, habits are reactive stimulus-response pairings that are strengthened when followed by reward. Executing a habit does not require planning toward a valued outcome, and thus correspond to the alternative class of model-free algorithms.

A key test for model-based versus model-free control is to assess whether a person continues to value an action even when it's connection to reward has been broken. A model-based system immediately devalues the action because it plays no productive role in maximizing expected outcomes, whereas a model-free learning system continues to assign value to the action based on its prior history of reward. In this sense, model-free algorithms assign value directly to actions, whereas model-based algorithms assign value to outcomes and then derive action values via online planning.

Many moral norms exhibit this signature property of model-free valuation. For instance, some American travelers feel compelled to tip foreign waiters 20% even when there is no such local norm. Presumably this does not reflect an underlying concern for the relevant outcome (well-funded foreign waitstaffs), but rather the habit-like internalization of an action-based value: Good service requires a tip. Indeed, evidence suggests that such altruistic actions are supported by internalized norms deployed automatically [28]. Likewise, in the trolley problem an outcome-based assessment favors doing direct harm to a focal individual, but people find it difficult to endorse such harm. This can be understood as the consequence of negative value assigned intrinsically to an action: direct, physical harm [29].

Research on habit learning has centered largely on the computational role of dopaminergic targets in the basal ganglia. Current neuropsychological research provides little association, however, between abnormal moral behavior and insult to the basal ganglia. Moreover, motor habits triggered by the basal ganglia are typically not accompanied by the subjective experience of value in the way that morals are: Tying your shoes feels automatic, but not desperately important. A more likely candidate for the encoding of action-based moral values is the ventromedial prefrontal cortex (vmPFC) [30–32]. As such, a key area for future research is to assess the role of model-free value representation in vmPFC [33•], especially in the moral domain. Also, while some moral values include