The Cognitive Neuroscience of Moral Judgment and Decision-Making

JOSHUA D. GREENE AND LIANE YOUNG

ABSTRACT This article reviews recent history and advances in the cognitive neuroscience of moral judgment and behavior. This field is conceived not as the study of a distinct set of neural functions but as an attempt to understand how the brain’s core neural systems coordinate to solve problems that we define, for nonneuroscientific reasons, as “moral.” At the heart of moral cognition are representations of value and the ways in which they are encoded, acquired, and modulated. Research dissociates distinct value representations—often within a dual-process framework—and explores the ways in which representations of value are informed or modulated by knowledge of mental states, explicit decision rules, the imagination of distal events, and social cues. Studies illustrating these themes examine the brains of morally pathological individuals, the responses of healthy brains to prototypically immoral actions, and the brain’s responses to more complex philosophical and economic dilemmas.

Cognitive neuroscience aims to understand the mind in physical terms. Against this philosophical backdrop, the cognitive neuroscience of moral judgment takes on special significance. Moral judgment is, for many, the quintessential operation of the mind beyond the body, the earthly signature of the soul. Indeed, in many religious traditions it’s the quality of a soul’s moral judgment that determines where it ends up. Thus, the prospect of understanding morality in physical terms may be especially alluring, or unsettling, depending on your point of view. In this brief review we provide a progress report on these efforts. Here we focus on research using neuroscientific/biological methods, but we regard this as an artificial restriction, useful only for limiting our scope.

The Paradox of the “Moral Brain”

The fundamental problem with the “moral brain” is that it threatens to take over the entire brain and thus ceases to be a meaningful neuroscientific topic. This is not because morality is meaningless but rather because neuroscience is centrally concerned with physical mechanisms, and it’s increasingly clear that morality has few, if any, neural mechanisms of its own (Young & Dun- gan, 2012). By way of analogy, the things we call vehicles are bound together, not by their internal mechanics—which include, pedals, sails, and nuclear reactors—but by their common function. So, too, with morality. More specifically, we regard morality as a suite of cognitive mechanisms that enable otherwise selfish individuals to reap the benefits of cooperation (Frank, 1988; Greene, 2013). Humans have psychological features that are straightforwardly moral (such as empathy) and others that are not (such as in-group favoritism) because they enable us to achieve goals that we can’t achieve through pure selfishness. We won’t defend this controversial thesis here. Instead, our point is that if this unified theory of morality is correct, it doesn’t bode well for a unified theory of moral neuroscience. Previously, some hoped to find a dedicated “moral organ” in the brain (Hauser, 2006). It’s now clear, however, that the “moral brain” is, more or less, the whole brain, applying its computational powers to problems that we, for nonneuroscientific reasons, classify as “moral.”

Understanding this is, itself, a kind of progress, but it leaves the cognitive neuroscience of morality—and the authors of a chapter that would summarize it—in an awkward position. To truly understand the neuroscience of morality, we must understand the many neural systems that shape moral thinking, none of which, so far, appears to be specifically moral. At the heart of moral cognition are interlocking systems that represent the value of actions and outcomes (Bartra, McGuire, & Kable, 2013; Craig, 2009; Knutson, Taylor, Kaufman, Peterson, & Glover, 2005). Representations of value are informed and modulated by systems that represent mental states (Frith & Frith, 2006; Koster-Hale et al., 2017) and that orchestrate thought and action in accordance with more abstract knowledge, rules, and goals (Miller & Cohen, 2001). This often gives rise to a dual-process dynamic, whereby automatic processes compete with more controlled processes (Kahneman, 2003).
Other systems enable us to imagine complex distal events (Buckner, Andrews-Hanna, & Schacter, 2008) and keep track of who’s who in the social world (Gikara & Van Bavel, 2014). These computational themes recur in lessons learned from abnormally antisocial brains, the responses of healthy brains to basic transgressions, and the ways in which our brains resolve more complex philosophical and economic dilemmas.

**Bad Brains**

The neuroscience of morality began with the study of brain damage leading to antisocial behavior. Such research accelerated in the 1990s with a series of path-breaking studies of decision-making in patients with damage to ventromedial prefrontal cortex (vmPFC), one of the regions damaged in the famous case of Phineas Gage (Damasio, 1994). Such patients made poor real-life decisions, but their deficits typically evaded detection using conventional measures of executive function (Saver & Damasio, 1991) and moral reasoning (Anderson, Bechara, Damasio, Tranel, & Damasio, 1999). Using a game designed to simulate real-world risky decision-making (the Iowa Gambling Task), Bechara, Tranel, Damasio, and Damasio (1996) documented these behavioral deficits and demonstrated, using autonomic measures, that these deficits are emotional. It seems that such patients make poor decisions because they lack the feelings that guide complex decision-making in healthy individuals. These early studies identified the vmPFC as critical for affectively driven moral choice and underscored the role of learning in moral development, as early-onset vmPFC damage leads not only to poor judgment but to a more psychopathic behavioral profile (Anderson et al., 1999).

Psychopathy is characterized by a pathological degree of callousness, a lack of empathy or emotional depth, a lack of genuine remorse for antisocial actions (Hare, 1991), and a tendency toward instrumental aggression (Blair, 2001). Psychopaths exhibit profound emotional deficits. In clinical and subclinical psychopathy, the amygdala, which plays a central role in emotional learning and memory (Phelps, 2006), exhibits weaker responses to fearful faces (Marsh et al., 2008) and to depictions of moral transgressions (Harenski, Harenski, Shane, & Kiehl, 2010). Critically, these muted affective responses are selective, responding to threats but not distress (Blair, Jones, Clark, & Smith, 1997). This pattern reemerges in more recent work showing that psychopaths, when prompted to imagine painful injuries to themselves and others, exhibit normal neural responses to their own imagined pain but reduced responses in the amygdala and insula, as well as reduced connectivity with the orbitofrontal cortex (OFC) and vmPFC, when imagining the pain of others (Decety, Skelly, & Kiehl, 2013). Likewise, a study of incarcerated psychopaths revealed reduced responses to distress cues in the vmPFC/OFC (Decety, Skelly, & Kiehl, 2013). A similar pattern, featuring the amygdala, has been observed in youths with psychopathic traits (Marsh et al., 2008, 2013).

Consistent with the above, Blair (2007) has proposed that psychopathy arises primarily from dysfunction in the amygdala, which is crucial for stimulus-reinforcement learning (Davis & Whalen, 2001). He argues further that psychopathy involves core deficits in response-outcome learning, which depends critically on the frontostriatal pathway, including the dorsal and ventral striatum as well as the vmPFC (Blair, 2017). This leads to abnormal socialization, such that psychopathic individuals fail to attach negative affective values to socially harmful outcomes and actions. These learning deficits manifest in judgment as well as behavior, such that psychopaths (or a subset thereof: Aharoni, Sinnott-Armstrong, & Kiehl, 2012) fail to distinguish between rules that authorities cannot legitimately change (“moral” rules—e.g., a classroom rule against hitting) from rules that authorities can legitimately change (“conventional” rules—e.g., a rule prohibiting talking out of turn; Blair, 1995).

Psychopaths, in addition to their weak affective responses to harm, tend to be impulsive (Hare, 1991). Psychopaths, compared to other incarcerated criminals, exhibit signs of reduced response conflict when behaving dishonestly (Abe, Greene, & Kiehl, 2018), and related responses to an impulse-control task (go/no-go) predict criminal rearrest (Aharoni et al., 2013). These deficits may ultimately derive from abnormal reward processing: psychopaths who harm impulsively exhibit heightened responses to reward within the frontostriatal pathway (Buckholtz et al., 2010).

An illuminating recent study (Darby et al., 2017) combines lesion data and resting-state functional connectivity data to explain why so many neural regions are implicated in antisocial behavior and why some of these regions appear to be more central than others. They find that the regions most reliably implicated in antisocial behavior are positively functionally connected to the frontostriatal pathway and/or the amygdala/anterior temporal lobe. By contrast, these regions tend to be negatively functionally connected to the frontoparietal control network, consistent with a dual-process framework (see below).

**Responsive Brains**

Consistent with studies of psychopathology, research on how healthy brains respond to moral transgressions
and opportunities highlights the importance of the frontostriatal pathway (Decety & Porges, 2011; Moll et al., 2006; Shenhav & Greene, 2010) and the amygdala-vmPFC circuit (Blair, 2007; Decety & Porges, 2011). Bookending their research in psychopaths, Marsh et al. (2014) have shown that extraordinary altruists (who have donated kidneys to strangers) tend to have larger amygdalae that are more sensitive to facial fear expressions. Likewise, several studies highlight the importance of the insula, which represents subjective value and appears to be an expanded somatosensory region (Craig, 2009). The insula’s responses reflect the aversiveness of moral transgressions (Baumgartner, Fischbacher, Feierabend, Lutz, & Fehr, 2009; Schaich Borg, Lieberman, & Kiehl, 2008), employing a multimodal code that also reflects pain, vicarious pain, disgust, and unfairness (Corradi-Dell’Acqua, Tusche, Vuilleumier, & Singer, 2016).

As Oliver Wendell Holmes Jr. famously observed, even a dog knows the difference between being tripped over and being kicked. Likewise, the human amygdala distinguishes between depictions of intentional and accidental harm within 200 ms, as revealed by depth electrode recordings (Hesse et al., 2016). The temporoparietal junction (TPJ) is the region most reliably implicated in the representation of morally relevant mental states and mental states more generally (Frith & Frith, 2006). The TPJ is especially sensitive to attempted harms (Koster-Hale, Saxe, Dungan, & Young, 2013; Young, Cushman, Hauser, & Saxe, 2007), which are wrong only because of the agent’s mental state. More recent evidence indicates that the TPJ separately encodes information about agents’ beliefs and values (Koster-Hale et al., 2017).

Both attempted harms and accidental harms set up a tension between outcome-based and intention-based judgment. This can give rise to a dual-process dynamic (see below), such that an understanding of mental states overrides an impulse to blame, or generates a more abstract reason to blame, despite the absence of harm. Consistent with this, TMS applied to the TPJ results in a childlike (Piaget, 1965), “no harm, no foul” pattern of judgment in which attempted harms are judged less harshly (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). In addition, a network of brain regions, including the TPJ and dorsal anterior cingulate cortex (ACC), appear to suppress amygdala responses to emotionally salient unintentional transgressions (Treadway et al., 2014). The “no harm, no foul” pattern is also observed in patients with vmPFC damage (Young, Bechara, et al., 2010), connecting the aforementioned effects in the amygdala and TPJ to the frontostriatal pathway. Consistent with this, psychopaths (Young, Koenigs, Kruepke, & Newman, 2012) and patients with alexithymia (Patil & Salani, 2014a), a condition that reduces awareness of one’s own emotional states, judge accidental harms to be more acceptable, reflecting reduced affective responses to harmful outcomes. Individuals with high-functioning autism exhibit a complementary pattern, “if harm, then foul,” judging accidental harms unusually harshly (Moran et al., 2011). Finally, split-brain patients (Miller et al., 2010), like vmPFC patients, exhibit a “no harm, no foul” pattern, indicating that sensitivity to intention depends on the integration of information across the cerebral hemispheres.

**Puzzled Brains**

To better understand more complex moral judgments, researchers have used moral dilemmas that capture the tension between competing moral considerations. The research described above emphasizes the role of emotion (Haidt, 2001), while traditional developmental theories emphasize controlled reasoning (Kohlberg, 1969). Greene and colleagues (Greene, 2013; Greene et al., 2001, 2004) have developed a dual-process (Kahneman, 2003) theory of moral judgment that synthesizes these perspectives. More specifically, this theory associates controlled cognition with utilitarian/consequentialist moral judgment aimed at promoting the greater good (Mill, 1861/1998) while associating automatic emotional responses with competing deontological judgments that are naturally justified in terms of rights or duties (Kant, 1785/1959).

This theory was inspired by a long-standing philosophical puzzle known as the *trolley problem* (Foot, 1978; Thomson, 1985). In the *switch* version of the problem, one can save five people who are mortally threatened by a runaway trolley by hitting a switch that will turn the trolley onto a side track, killing one person. Here, most people approve of acting to save more lives. In the contrasting *footbridge* dilemma, the only way to save the five is to push a large person off a footbridge and into the trolley’s path. Here, most people disapprove. Why the difference? And what does this tell us about moral judgment?

In short, people say no to the action in the *footbridge* case because that action elicits a relatively strong negative emotional response, and this response tends to override the cost-benefit reasoning that favors pushing. In the *switch* case, the harmful action is less emotionally salient, and therefore cost-benefit reasoning tends to prevail. The first evidence for these conclusions came from a functional magnetic resonance imaging (fMRI) study (Greene et al., 2001) that contrasted sets of “personal” and “impersonal” dilemmas loosely modeled after the *footbridge* and *switch* cases. It found that
“personal” dilemmas elicited increased activity in the mPFC, medial parietal cortex, and TPJ. These regions were previously associated with emotion and are now recognized as comprising most of the default mode network (DMN) (Buckner, Andrews-Hanna, & Schacter, 2008). In contrast, the “impersonal” dilemmas elicited relatively greater activity in the frontoparietal control network. A subsequent experiment found increased activity for utilitarian judgment within this network, including regions of DLPFC (Greene et al., 2004). Likewise, a more recent study found increased engagement of the DLPFC when participants were instructed to focus exclusively on utilitarian outcomes (Shenhav & Greene, 2014). Greene et al. (2004) also found increased amygdala responses to “personal” dilemmas. More recent evidence indicates that the DMN’s response to “personal” dilemmas is best understood not as an emotional response per se but as the increased engagement of a mechanism that enables the construction and representation of nonpresent episodes such as memories of the past, “prospections” of the future, and hypothetical imaginings (Buckner, Andrews-Hanna, & Schacter, 2008; DeBrigard, Addis, Ford, Schacter, & Giovanelli, 2013). Consistent with this, Amit and Greene (2012) found that individuals with more visual cognitive styles tend to make fewer utilitarian judgments in response to high-conflict personal dilemmas and that disrupting visual imagery while contemplating these dilemmas increases utilitarian judgment.

Some of the most compelling evidence for the dual-process theory comes from studies of patients with emotion-related deficits. Mendez, Anderson, and Shapiro (2005) found that patients with frontotemporal dementia, who are known for their “emotional blunting,” are disproportionately likely to approve of the utilitarian action in the footbridge dilemma. Likewise, patients with vmPFC lesions make up to five times as many utilitarian judgments in response to standard high-conflict dilemmas (Ciaramelli, Muccioli, Ladávás, & di Pellegrino, 2007; Koenigs et al., 2007). Such patients also make more utilitarian judgments in response to dilemmas pitting familial duty against the greater good (e.g., your sister vs. five strangers; Thomas, Croft, & Tranell, 2011). As expected, vmPFC patients exhibit correspondingly weak physiological responses when making utilitarian judgments (Moretto, Ladávás, Mattioli, & di Pellegrino, 2010), and healthy people who are more physiologically reactive are less likely to make utilitarian judgments (Cushman, Gray, Gaffey, & Mendes, 2012). Paralleling their more lenient responses to accidental harms (see above), low-anxiety psychopaths (Koenigs et al., 2012) and people with alexithymia (Koven, 2011; Patil & Silani, 2014b) are also more approving of utilitarian sacrifices. Critically, these effects depend not only on the disruption of the affective pathway that favors deontological judgment but also on a preserved capacity for cost-benefit reasoning, without which their judgments would simply be disordered, rather than more utilitarian.

Other studies using dilemmas highlight the shared and distinctive functions of the amygdala and vmPFC. Citalopram—a selective serotonin-reuptake inhibitor (SSRI) that increases emotional reactivity in the short term through its influence on the amygdala and vmPFC—increases deontological judgment (Crockett, Clark, Hauser, & Robbins, 2010). By contrast, lorazepam, an antianxiety drug, has the opposite effect (Perkins et al., 2012), as does the administration of testosterone (Chen, Decety, Huang, Chen, & Cheng, 2016). Consistent with this, individuals with psychopathic traits exhibit reduced amygdala responses to personal moral dilemmas (Glenn, Raine, & Schug, 2009). In healthy people, amygdala activity tracks self-reported emotional responses to harmful transgressions and predicts deontological judgments in response to them (Shenhav & Greene, 2014). The same study shows a different pattern for the vmPFC, which is most active when people have to make integrative, “all things considered” judgments, as compared to simply reporting on emotional reactions or assessing options solely in terms of their consequences. This suggests that the amygdala generates an initial negative response to personally harmful actions while the vmPFC weighs that signal against a competing signal reflecting the value of the greater good (see also Hutcherson, Montaser-Koushars, Woodward, & Rangel, 2015).

The vmPFC (along with the ventral striatum) also represents expected moral value, integrating information concerning the number of lives to be saved and the probability of saving them (Shenhav & Greene, 2010). These findings are consistent with our understanding of the frontostriatal pathway, and the vmPFC more specifically, as a domain-general integrator of decision values (Bartra, McGuire, & Kable, 2013; Knutson et al., 2005). We note that these structures evolved in mammals to evaluate goods, such as food, that tend to exhibit diminishing marginal returns. (The more food you’ve eaten, the less you need additional food.) This may explain our puzzling tendency to regard the saving of human lives as exhibiting diminishing marginal returns, as if the 100th life to be saved is somehow worth less than the first (Dickert, Västfjäll, Kleber, & Slovic, 2012).

Patients with hippocampal damage, unlike vmPFC patients, are less likely to make utilitarian judgments (McCormick, Rosenthal, Miller, & McGuire, 2016). This result is surprising (cf., Amit & Greene, 2012; Greene et al., 2001) but ultimately consistent with the
dual-process theory. The hippocampus is a critical node within the DMN (Buckner, Andrews-Hanna, & Schacter, 2008), which is, once again, essential for the imagination of nonpresent events. The inability of hippocampal patients to fully imagine dilemma scenarios may thus cause them to rely more on emotional responses to the types of actions proposed, as reflected in skin-conductance responses and self-reports (for contrasting null results, however, see Craver et al., 2016).

In an important theoretical development, Cushman (2013) and Crockett (2013) have proposed that the dissociation between deontological and utilitarian/consequentialist judgment reflects a more general dissociation between model-free and model-based learning systems (Daw & Doya, 2006). Model-free learning mechanisms assign values directly to actions based on past experience, while model-based learning attaches values to actions indirectly by attaching values to outcomes and linking outcomes to actions via internal models of causal relations. Thus, an action may seem wrong “in itself” because past experience has associated actions of that type (e.g., pushing people) with negative consequences (e.g., social disapproval), and yet the same action may seem right because it will, according to one’s causal world model, produce optimal consequences (saving five lives instead of one). Thus, the fundamental tension in normative ethics, reflected in the competing philosophies of Kant and Mill, may find its origins in a competition between distinct, domain-general mechanisms for assigning values to actions. With respect to the more deontological judgments made by hippocampal patients, McCormick et al. (2016) suggest that their judgments, influenced by a limited capacity for imagination, may be understood as relatively model-free.

Trolley dilemmas are, perhaps, an unlikely tool for scientists, and some researchers have questioned their widespread use. Kahane et al. (2015) have claimed that the utilitarian judgments they elicit are not truly utilitarian and merely reflect antisocial tendencies. This critique is based largely on a misunderstanding about how the term utilitarian has been used. The judgments are called utilitarian because they are required by utilitarianism and are thought to reflect simple cost-benefit reasoning, not because the judges are thought to be generally committed to utilitarian values (Conway, Goldstein-Greenwood, Polacek, & Greene, 2018). (One can make a utilitarian judgment without being a utilitarian, just as one can make an Italian meal without being Italian.) Addressing the provocative claim that utilitarian judgments are motivated entirely by antisocial tendencies, a series of studies replicating Kahane et al.’s studies with the addition of process dissociation measures confirms the predictions of the dual-process theory: utilitarian judgments reflect both decreased concern about causing harm and increased concern for the greater good (Conway et al., 2018). Conway et al. also examined the judgments of professional philosophers and showed, contra Kahane (2015), that trolley judgments do indeed reflect the fundamental tension between consequentialists and deontologists. Others have challenged the use of hypothetical dilemmas based on concerns about their ecological validity (e.g., Bostyn, Sevenhant, & Roets, 2018). For replies, see Conway et al. (2018) and Plunkett and Greene (in press).

**Cooperative Brains**

Research on altruism and cooperation, though often considered apart from “morality,” could not be more central to our understanding of the moral brain. The most basic question about the cognitive neuroscience of altruism and cooperation is this: What neural processes enable and motivate people to be “nice”—that is, to pay costs to benefit others?

Consistent with our evolving story, the value of helping others, in both unidirectional altruism and bidirectional cooperation, is represented in the frontostriatal pathway and modulated by both economic incentives and social signals (Declerck, Boone, & Emonds, 2013). Activity in this pathway tracks the value of charitable contributions (Moll et al., 2006) and of sharing resources with other individuals (Zaki & Mitchell, 2011). Likewise, it encodes the discounted value of rewards gained at the expense of others (Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017). Here, signals from the DLPFC appear to modulate striatal signals, resulting in more altruistic behavior. The same pattern is observed in the case of increased altruism following compassion training (Weng et al., 2013). Striatal signals, likewise, track the value of punishing transgressors (Crockett et al., 2013; de Quervain et al., 2004; Singer et al., 2006). And, as above, the DMN appears to have a hand in altruism: TPJ volume (Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012) and medial PFC activity (Waytz, Zaki, & Mitchell, 2012) both predict altruistic behavior, with more dorsal mPFC regions representing the value of rewards for others (Apps & Rammnani, 2014).

As noted above, the brain uses its endogenous carrots—reward signals—to motivate cooperative behavior. It also uses its sticks—negative affective responses to uncooperative behavior. Activity in the insula scales with the unfairness of ultimatum game (UG) offers (Gabay, Radua, Kempton, & Mehta, 2014; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003) including offers to third parties (Corradi-Dell’Acqua, Civai, Rumiai, & Fink, 2012). Insula responses also predict aversion to
inequality in the distribution of resources (Hsu, Anen, & Quartz, 2008) and egalitarian behavior and attitudes (Dawes et al., 2012). The insula and the amygdala both respond to the punishment of well-behaved people (Singer, Kiebel, Winston, Dolan, & Frith, 2004). Perhaps surprisingly, vmPFC damage leads to increased rejection of unfair UG offers (Koenigs & Tranel, 2007), mirroring patterns observed in psychopaths (Koenigs, Kruepke, & Newman, 2010.) This may be because the vmPFC integrates signals responding to material gain as well as unfairness (which compete in the UG) and because, in the absence of such signals, one applies a reciprocity rule.

Honesty is a form of cooperation, and dishonesty is a form of defection. Greene and Paxton (2009) gave people repeated opportunities to gain money by lying about their accuracy in predicting the outcomes of coin flips. Consistently honest subjects appeared to be “gracefully” honest, exhibiting no additional engagement of the frontoparietal control network in forgoing dishonest gains. By contrast, subjects who behaved dishonestly exhibited increased control-related activity, both when lying and when refraining from lying. These individual differences in (dis)honesty are predicted by striatal responses to rewards in an unrelated task (Abe & Greene, 2014). Baumgartner et al. (2009) describe a similar dual-process dynamic in which breaking promises involves increased engagement of the amygdala and the frontoparietal control network.

Cooperation depends on trust, which in turn requires evaluating people’s trustworthiness (Delgado, Frank, & Phelps, 2005). We describe the people we trust as “close,” and this metaphor is reflected in how the brain represents social relationships: A region of the inferior parietal lobe has been shown to represent spatial, temporal, and social proximity using a common code, as demonstrated by cross-trained pattern classification (Parkinson, Liu, & Wheatley, 2014). Cooperation is more likely with friends than strangers, and the additional social value of cooperation with friends is reflected in ventral-striatal signals and in the mPFC (Fareri, Chang, & Delgado, 2015). Likewise, our brains respond differently to in-group and out-group members, including members of “minimal” groups formed in the lab (Cikara & Van Bavel, 2014). Both neural and behavioral data indicate that cooperation with in-group members is rewarding and relatively effortless, while cooperation with out-group members engages more cognitive control (Hughes, Ambady, & Zaki, 2017, consistent with evolutionarily inspired theories of dual-process cooperation (Bear & Rand, 2016; Greene, 2013; Rand, Greene, & Nowak, 2012. But see Everett, Ingbretsen, Cushman, and Cikara (2017) for evidence of intuitive cooperation with “minimal” out-groups).

Oxytocin is a neuropeptide implicated in social attachment and affiliation across mammals (Insel & Young, 2001). In humans it’s been associated with empathy and prosocial behavior (Bartz et al., 2015; Heinrichs, von Dawans, & Domes, 2009). An early and influential study found that intranasally administered oxytocin increases trust among strangers (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005), and many studies have associated variation in the oxytocin receptor gene (OXTR) with morally relevant phenotypes, including empathic concern (Rodrigues, Saslow, Garcia, John, & Keltner, 2009), generosity (Israel et al., 2009), and psychopathy (Dadds et al., 2014). As with many candidate gene studies, subsequent studies with larger samples have failed to replicate many such effects (Apicella et al., 2010; Bakermans-Kranenburg & van IJzendoorn, 2014), and doubts have been raised about the relation between oxytocin and trust (Nave, Camerer, & McCullough, 2015). A recent study employing separate exploratory and confirmatory samples found an association between an OXTR variant and two types of dilemma judgments (Bernhard et al., 2016).

Recent research indicates that the effects of oxytocin are highly variable across personality types (Bartz et al., 2015) and sex (Rilling et al., 2014) and may even include antisocial behavior (Ne’eman, Perach-Barzilay, Fischer-Sholty, Atias, & Shamay-Tsoory, 2016). According to a recent influential theory, the variable effects of oxytocin across individuals, contexts, and relationships are best understood as effects of heightening the salience of social cues, again through modulation of the fronto-striatal pathway (Shamay-Tsoory & Abu-Akel, 2016). Most notable of all, there is mounting evidence that the effects of oxytocin are “parochial,” biasing judgment and behavior in favor of in-group members (De Dreu et al., 2010; Shalvi & De Dreu, 2014).

Although such results were surprising, given oxytocin’s well-established role in affiliative behavior, they make evolutionary sense. Morality evolved, not as a device for universal cooperation but as a competitive weapon—as a system for turning Me into Us, which in turn enables Us to outcompete Them. It does not follow from this, however, that we are doomed to be warring tribalists. Drawing on our ingenuity and flexibility, it’s possible to put human values ahead of evolutionary imperatives, as we do when we use birth control.

Looking Back, and Ahead

How does the moral brain work? Answer: exactly the way you’d expect it to work if you understand (1) which
cognitive functions morality requires and (2) which cognitive functions are performed by the brain’s core neural systems. Our conclusion that human morality depends on the brain’s general-purpose machinery for representing value, applying cognitive control, mentalizing, reasoning, imagining, and reading social cues will come as no surprise to today’s neuroscientists. But the emergence of morality as a source of tractable neuroscientific problems is itself significant. For the broader sciences and the general public, our increasingly detailed, mechanistic understanding of human morality is radically demystifying, challenging traditional dualistic assumptions about human nature with important implications for law, public policy, and our collective self-image (Farah, 2012; Greene & Cohen, 2004; Shariff et al., 2014).

From its inception, cognitive neuroscience has focused on structure-function relationships, teaching us which parts of the brain do what. By contrast, we know very little about how ideas move around and interact in the brain. We can track our neural responses to the thought of pushing someone off of a footbridge, but how do our brains even compose such a thought in the first place? We are just beginning to understand how the brain can represent, for example, the morally significant difference between a baby kicking a grandfather and a grandfather kicking a baby (Frankland & Greene, 2015)—a modest step. However, with the confluence of multivariate analysis methods (Kriegeskorte, Goebel, & Bandettini, 2006; Norman, Polyn, Detre, & Haxby, 2006), network approaches (Bullmore & Sporns, 2009), and neurally inspired models of high-level cognition (Graves et al., 2016; Kriete et al., 2013; Lake, Ullman, Tenenbaum, & Gershman, 2017), we may finally be ready to understand how the brain flexibly and precisely manipulates the contents of thoughts (Fodor, 1975; Marcus, 2001). And that’s a good thing, because understanding moral thinking may require a more general understanding of thinking.

Acknowledgments

Many thanks to Catherine Holland for research assistance. Thanks to Joshua Buckholz, Joe Paxton, Adina Roskies, and Walter Sinnott-Armstrong for helpful comments.

REFERENCES


Amit, E., & Greene, J. D. (2012). You see, the ends don’t justify the means: Visual imagery and moral judgment. Psychological Science, 23(8), 861–868.


Everett, J. A., Inghrismen, Z., Cushman, F., & Cikara, M. (2017). Deliberation erodes cooperative behavior—even towards competitive out-groups, even when using a control
condition, and even when eliminating selection bias. *Journal of Experimental Social Psychology, 73,* 76–81.


Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131.


---1

---0

---+1


1014 NEUROSCIENCE AND SOCIETY


