CHAPTER 1

# Can We Understand Moral Thinking without Understanding Thinking?

Joshua D. Greene



**Can we understand moral thinking without understanding thinking?**

Only up to a point; to understand morality well enough to put it into a flexibly behaving machine, we must first learn more about how our brains compose and manipulate structured thoughts.

Nerds of a certain age will recall Commander Data from *Star Trek: The Next Generation*, the humanoid android on a personal quest to become more human. Data's positronic brain features an "ethical subroutine," a computational add-on designed to enhance his capacity for moral judgment. The field of moral cognition has bad news for Commander Data. His ethical subroutine may be wonderful, but it's not making him more human.

As far as we can tell, there is nothing in our brains specifically dedicated to moral thinking (Greene, 2014; Parkinson et al., 2011; Ruff & Fehr, 2014; Young & Dungan, 2012). (But see Hauser, 2006, and Mikhail, 2011, for a dissenting view). Observe human brains engaged in moral judgment and you'll see neural activity representing the values of available alternatives (Blair, 2007; Hutcherson et al., 2015; Moll et al., 2006; Shenhav & Greene, 2010, 2014; Zaki & Mitchell, 2011; Hutcherson et al., 2015), explicit decision rules (Greene et al., 2004; Greene & Paxton, 2009), structured behavioral events (Frankland & Greene, 2015), and people's intentions (Young, Cushman, Hauser, & Saxe, 2007; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). Critically, these neural pathways, when engaged in moral cognition, appear to be doing the same things they do in other contexts that have nothing in particular to do with morality, such as making trade-offs between risk and reward (Knutson, Taylor, Kaufman, Peterson, & Glover, 2005), overriding automatic responses based on explicit task demands (Miller & Cohen, 2001), imagining distal events (Buckner, Andrews-Hanna, & Schacter, 2008; De Brigard, Addis, Ford, Schacter, & Giovanello, 2013), understanding who did what to whom (Wu, Waller, & Chatterjee, 2007), and keeping track of who believes what (Mitchell, 2009; Saxe, Carey, & Kanwisher, 2004). It's not just that neuroscientific data are too coarse-grained to distinguish the distinctively moral patterns of thinking from the rest. Behavioral stud-

ies indicate that moral and nonmoral thinking follow similar patterns and make use of shared computational resources when we evaluate options (Crockett, 2013, 2016; Cushman, 2013; Krajbich, Hare, Bartling, Morishima, & Fehr, 2015), reason (Paxton, Ungar, & Greene, 2012), imagine (Amit & Greene, 2012), and understand the minds of others (Moran et al., 2011). Cognitively speaking, morality does not appear to be special.

If morality isn't "a thing" in the brain, then what exactly are researchers who specialize in moral psychology trying to understand? I believe that morality can be a meaningful scientific topic even if moral cognition has no distinctive cognitive mechanisms of its own. An analogy: Motorcycles and sailboats have very little in common at the mechanistic level, respectively resembling nonvehicles such as lawn mowers and kites more than they resemble each other. Nevertheless, they are both vehicles in good standing. They rightly belong to the same category because of what they do, not how they do it. In the same way, the various kinds of thinking we call moral may be bound together, not by their engagement of distinctive cognitive mechanisms but by the common function they serve: enabling otherwise selfish individuals to reap the benefits of social existence (Frank, 1988; Gintis, 2005; Greene, 2013; Haidt, 2012). If this functional account of morality is correct, then moral cognition, as a field or subfield, is best understood as a bridge. It's an attempt to connect the concepts of everyday moral life—right and wrong, good and bad, virtue and vice—to the subpersonal mechanisms of the mind and brain. Bridges are exciting to build and useful once completed, but they are rarely destinations of their own. What happens after the bridge opens? Where does the traffic go?

On the neuroscientific side, the field of moral cognition has focused on implicating rather general cognitive functions and corresponding neural regions and networks. For example, there has been some debate concerning the relative roles of intuitive and affective processes on the one hand and more controlled, rule-based reasoning on the other (Greene, 2013; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001, 2012; Kohlberg, 1969; Turiel, 2006). This debate has featured evidence implicating brain regions associated primarily with emotion (Ciaramelli, Muccioli, Ladàvas, & di Pellegrino, 2007; Koenigs et al., 2007; Shenhav & Greene, 2014), along with other brain regions associated primarily with cognitive control (Cushman, Murray, Gordon-McKeon, Wharton, & Greene, 2012; Greene et al., 2004; Paxton & Greene, 2009; Shenhav & Greene, 2014). More recently, this contrast has been recast in terms of more basic computational principles (Crockett, 2013; Cushman, 2013), a welcome development. But in nearly all of our attempts to explain moral judgment and behavior in terms of neural mechanisms, the explanations have featured very general processes, not detailed content. For example, we may explain people's responses to the classic *footbridge* dilemma (Thomson, 1985) in terms of affective responses enabled by the amygdala and the ventromedial prefrontal cortex (vmPFC), along with a competing cost–benefit decision rule supported by the dorsolateral prefrontal cortex (DLPFC), but nowhere in the neural data is there anything specifically related to a trolley, train tracks, a footbridge, pushing one person, or saving the lives of five. We know this information is in there, but we've only the most coarse-grained theories about how these details are represented and transformed in the process of moral judgment.

In behavioral research, detailed content plays a more prominent role. We distinguish between different ways of causing harm (Cushman, Young, & Hauser, 2006; Greene et al., 2009; Spranca, Minsk, & Baron, 1991), different kinds of moral violations and norms (Graham et al., 2011; Janoff-Bulman, Sheikh, & Hepp, 2009; Young & Saxe, 2011), different moral roles (Gray & Wegner, 2009), and much more besides. But these content-based distinctions and effects, however interesting and useful they may be, seem more like hints—intriguing products of the underlying cognitive mechanisms, rather than descriptions of those mechanisms. If Commander Data ever learns to think about moral questions like a human, he'll be sensitive to the act/omission distinction, care less about people's intentions when they do things that are disgusting, and so on. But we

currently have no idea how we would actually program or train in these features.

The problem, I believe, is that we're trying to understand moral thinking in the absence of a more general understanding of thinking. When you hear about a moral dilemma, involving, say, a trolley headed for five unsuspecting people and a footbridge, your brain responds to this string of words by activating a set of conceptual representations (TROLLEY, FOOTBRIDGE, FIVE, MAN, etc.). These representations are not merely activated to form a semantic stew of trolleyness, footbridgeness, and so forth. Rather, they are combined in a precise way to yield a highly specific structured representation of the situation in question, such that it's the five on the tracks, the man on the footbridge, the trolley headed toward the five, and you with the option to push the man in the name of the greater good. What's more, our brains naturally construct a representation of the situation so as to fill in countless unstated facts, such as the fact that the man, if pushed, will fall quickly through the air rather than gently floating to the ground like a feather. Our understanding of how all of this cognitive infrastructure works is rather limited. In saying this, I do not mean to discount the great strides made by philosophers (e.g., Fodor, 1975; Frege, 1976), linguists (e.g., Fillmore, 1982; Talmy, 2000), psychologists (e.g., Johnson-Laird, 1983; Johnson-Laird, Khemlani, & Goodwin, 2015; Kriete, Noelle, Cohen, & O'Reilly, 2013; Marcus, 2001; Pinker, 1994, 2007), and neuroscientists (e.g., Fedorenko, Behr, & Kanwisher, 2011; Friederici et al., 2003; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Pallier, Devauchell, & Dehaene, 2011) in addressing this large problem. What I mean is that we still lack a systematic understanding of what David Hume (1739/1978) and other Enlightenment philosophers called "the Understanding" and what Fodor (1975) called the "language of thought."

How well can we understand moral thinking—or any other kind of high-level thinking—without understanding the underlying mechanics of thought? Pretty well, some might say. This worry about underlying mechanisms could just be fetishistic reductionism. If "really" understanding moral thinking requires deciphering the language of thought, why stop there? To "really" understand the language of thought, don't we need to understand how populations of neurons represent things more generally? And to "really" understand that, don't we need a better of understanding of neurophysiology? And beneath that, must we not understand organic chemistry, chemical physics, and so on? Does this not lead to the absurd conclusion that the only "real" understanding of anything comes from particle physics?

I sympathize with this objection, but I think it goes too far. How far down the reductionist hierarchy we must go depends on what we're trying to do and what we get for our deeper digging. If you're a sailor, you need to understand the weather, but understanding the physics and chemistry of the atmosphere probably won't do you much additional good. By contrast, if you're developing models of weather and climate, pushing the bounds of long-range prediction, a detailed knowledge of the underlying mechanics is surely essential. Today, much of psychology, including moral psychology, looks more like sailing than cutting-edge atmospheric modeling. We isolate a specific variable in a specific and somewhat artificial context, and, if all goes well, we can say something about the general direction and size of the effect of manipulating that variable in that context. But if our long-term goal is to understand and predict real human behavior in complex circumstances, with many behaviorally significant variables operating simultaneously, we'll probably have to understand the thinking behind that behavior in a more encompassing way, not just in terms of "effects" but in terms of the underlying cognitive causes of those effects. I doubt that we'll need to descend into particle physics, but I suspect that we'll have to go significantly deeper than we currently do. In the best case, we'll understand the infrastructure of high-level cognition in sufficient detail that we could program or train Commander Data to think as we do—morally and otherwise.

Following this hunch, I and my collaborators have begun to pursue more basic questions about the nature of high-level cognition and its neural basis: How does the brain combine concepts to form thoughts (Frank-

land & Greene, 2015)? How are thoughts manipulated in the process of reasoning? How do thoughts presented in words get translated into mental images? And how do our brains distinguish the things we believe from the things we desire or merely think about? I don't know whether these investigations will bear fruit for moral psychology, sometime soon or ever. But this kind of research seems to me worth pursuing for its own sake, and there's a chance that it will teach us things about morality that we can't learn any other way.

## 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.

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

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences, 1124*(1), 1–38.

Ciaramelli, E., Muccioli, M., Ladàvas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 2,* 84–92.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17*(8), 363–366.

Crockett, M. J. (2016). How formal models can illuminate mechanisms of moral judgment and decision making. *Current Directions in Psychological Science, 25*(2), 85–90.

Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2012). Judgment before principle: Engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience, 7*(8), 888–895.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia, 51*(12), 2401–2414.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the USA, 108*(39), 16428–16433.

Fillmore, C. (1982). Frame semantics. In The Linguistic Society of Korea (Ed.), *Linguistics in the morning calm* (pp. 111–137). Seoul: Hanshin.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Cambridge, MA: Harvard University Press.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. New York: Norton.

Frankland, S. M., & Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences of the USA, 112*(37), 11732–11737.

Frege, G. (1976). *Logische untersuchungen* (2nd suppl. ed.). Göttingen, Germany: Vandenhoeck und Ruprecht.

Friederici, A. D., Rueschemeyer, S. A., Hahne, A., & Fiebach, C. J. (2003). The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cerebral Cortex, 13*(2), 170–177.

Gintis, H. (Ed.). (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge, MA: MIT Press.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*(2), 366–385.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*(3), 505.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: Penguin.

Greene, J. D. (2014). The cognitive neuroscience of moral judgment and decision-making. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences V* (pp. 1013–1023). Cambridge, MA: MIT Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.

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

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.

Hagoort. P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438–441.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J. (2012). *The righteous mind: Why good people are divided by religion and politics*. New York: Pantheon.

Hauser, M. D. (2006). The liver and the moral organ. *Social Cognitive and Affective Neuroscience, 1*(3), 214–220.

Hume, D. (1978). *A treatise of human nature* (L. A. Selby-Bigge & P. H. Nidditch, Eds.). Oxford, UK: Oxford University Press. (Original work published 1739)

Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex. *Journal of Neuroscience, 35*(36), 12593–12605.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453–458.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology, 96*(3), 521–537.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness (No. 6)*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences, 19*(4), 201–214.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience, 25*(19), 4806–4812.

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.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Chicago: Rand McNally.

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A common mechanism underlying food choice and social decisions. *PLOS Computational Biology, 11*(10), e1004371.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013) Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences of the USA, 110*(41), 16390–16395.

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.

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

Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 364*(1521), 1309–1316.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the USA, 103,* 15623–15628.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences of the USA, 108*(7), 2688–2692.

Pallier, C., Devauchelle, A. D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences of the USA, 108*(6), 2522–2527.

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is morality unified?: Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience, 23*(10), 3162–3180.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177.

Pinker, S. (1994). *The language instinct*. New York: Harper Perennial Modern Classics.

Pinker, S. (2007). *The stuff of thought: Lan-*

*guage as a window into human nature*. New York: Viking.
Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15*(8), 549–562.
Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124.
Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron, 67*(4), 667–677.
Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience, 34*(13), 4741–4749.
Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*(1), 76–105.
Talmy, L. (2000). *Toward a cognitive semantics: Concept structuring systems* (Vols. 1–2). Cambridge, MA: MIT Press.
Thomson, J. (1985). The trolley problem. *Yale Law Journal, 94*, 1395–1415.
Turiel, E. (2006). Thought, emotions and social interactional processes in moral development. In M. Killen & J. Smetana (Eds.), *Handbook of moral development* (pp. 1–30). Mahwah, NJ: Erlbaum,
Wu, D. H., Waller, S., & Chatterjee, A. (2007). The functional neuroanatomy of thematic role and locative relational knowledge. *Journal of Cognitive Neuroscience, 19*(9), 1542–1555.
Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the USA, 107*(15), 6753–6758.
Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the USA, 104*(20), 8235–8240.
Young, L., & Dungan, J. (2012). Where in the brain is morality?: Everywhere and maybe nowhere. *Social Neuroscience, 7*(1), 1–10.
Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202–214.
Zaki, J., & Mitchell, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences of the USA, 108*(49), 19761–19766.