The Philosopher in the Theater[1]

Fiery Cushman Joshua Greene

Harvard University

The moral principles found in philosophy and embodied in law are often strikingly complex, peculiar, and yet resolutely persistent. For instance, it was long held in Britain that a person could be tried for murder only if the victim died within a year and a day of the crime. And in the United States, if a robber gets into a shootout with a cop and the cop's bullet hits a bystander, the robber can be charged with murdering the bystander. Naively, one might have assumed that murder could be defined simply as "intentionally causing another person to die." In fact, the American Law Institute's Model Penal Code requires pages of fine print.

Our goal in this chapter is to present a model of the origins of moral principles that explain these properties: complexity, peculiarity, and persistence. According to this model, abstract, general moral principles are constructed from the raw material of intuitive responses to particular cases, as explained in the next section of the chapter. Those intuitive responses depend, in turn, on many psychological capacities that are not specific to morality at all. These include attributions of causation ("John harmed Jane...") and intent ("… on purpose."). Consequently, explicit moral principles reflect the complexity of our psychological processes of causal and intentional attribution. That complexity can seem peculiar because our intuitive, automatic attributions of causation and intent are often at odds with our more considered, explicit theories of causation and intention. Despite their complexity and peculiarity, these principles persist because they are supported by compelling emotions that are particularly difficult to revise or reject, an issue considered in the final major section of this chapter.

Intuitions and Principled Reasoning

Ordinary peoples' moral judgments often track philosophers' and lawyers' explicit principles, mirroring their complexity and nuance. For a long while psychologists assumed a simple explanation of this phenomenon: Ordinary people use explicit principles when they make moral judgments (Kohlberg, 1969). (By an explicit moral principle, we mean a general moral rule that can be verbalized and is available to conscious reasoning.) However, recent research in moral psychology forcefully challenges this assumption (Haidt, 2001; see also Graham & Haidt, this volume). At least in some cases, people make moral judgments that are consistent with prominent philosophical or legal principles, and yet have no explicit awareness of those principles (Cushman, Young, & Hauser, 2006; Mikhail, 2000). In fact, some of these characteristic patterns of judgment are established by early childhood (Pellizzoni, Siegal, & Surian, 2010; see also Bloom, this volume). These findings indicate that our moral principles derive their content from our moral judgments, rather than the other way around.

To some extent, this should not come as a surprise. Many philosophers commonly test their principles against intuitions about specific cases, and some openly embrace the project of systematizing moral intuition in principles (Fischer & Ravizza, 1992; Kamm, 2006). Also, psychological research shows that ordinary people often attempt to provide post hoc rationalizations of their moral judgments, constructing explicit principles to match their intuitive responses to particular cases (Haidt, 2001; Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009). This suggests that the philosophical practice of constructing intuitively plausible moral principles may be continuous with the commonplace practice of rationalizing emotional moral commitments (Cushman & Young, 2009; Greene, 2007; Mikhail, 2000; Shweder & Haidt, 1993; see also Ditto & Liu, this volume).

From a certain perspective, this is a decidedly unflattering portrait of moral reasoning. For instance, pioneering experiments by Haidt showed that people are strikingly unable to provide adequate, consistent principles to support the common judgment that sibling incest is wrong (e.g., Haidt & Hersh, 2001). They claim it is wrong because it leads to genetic deficits in children; but they insist that it is still wrong

even if the sister is infertile. They claim it is wrong because the family would be embarrassed; but they insist that it is wrong even if kept secret. They claim it is wrong because the siblings will regret it; but they insist it is wrong even if the siblings find it quite enjoyable. Finally, they throw up their hands and say they do not know why incest wrong, they just know it is. Participants in these experiments appear to be engaged in a desperate and unsuccessful search for any principled basis for their intuitions, a phenomenon that Haidt calls "moral dumbfounding." It appears that their goal is not to develop a rational theory to guide future judgments, but instead to paint the veneer of reason over a foregone conclusion.

Yet, there is evidence that people's moral justifications are sometimes more considered and constrained than these early studies suggested. Take, for example, the tendency to justify the prohibition against incest by appeal to the possibility of harm (e.g., to the child, the family, or the siblings themselves; Paxton & Greene, in press). Consistent with these justifications, there is now mounting evidence that people engage in explicit utilitarian moral reasoning favoring actions that minimize harm (Bartels, 2008; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004). People's moral justifications also frequently invoke principles of causal responsibility (it is wrong to cause harm) and intent (it is wrong to harm another intentionally; see Cushman et al., 2006; Kohlberg, 1981; Piaget, 1965/1932), and again their judgments are generally consistent with these principles (e.g. Weiner, 1995).

There are other factors that play an important role in intuitive judgments, but which people nevertheless tend to regard as morally irrelevant. For instance, moral intuitions are sensitive to whether a perpetrator acts with direct physical force on a victim (Cushman et al., 2006; Greene et al., 2009): Pushing a person in front of a train is worse than flipping a switch that drops the person in front of a train. But when people recognize the role of force in their intuitive judgments, they often reject this factor as morally irrelevant (Cushman et al., 2006). For instance, one participant wrote, "I guess I was fooled by the line-pulling [i.e., indirect force] seeming more passive than the man-pushing, but that view is hard to justify now."

These studies suggest that, while moral intuitions play an important role in the development of explicit moral theories, people do not endorse intuitively supported moral principles indiscriminately. Rather, people seem to progress toward what philosophers call "reflective equilibrium," achieved through a tug-of-war between intuitive attitudes and principled commitments. If it is true that the philosopher's method of constructing moral principles is roughly continuous with the ordinary person's, then evidence for reflection is surely welcome news to philosophers. Maybe philosophers are experts in reflection: ordinary in the intuitions they harvest but extraordinary in their capacity to separate the wheat from the chaff.

A recent study tested this proposition by comparing intuitive judgments and moral principles in a large sample of philosophers — about 280 individuals who had earned a Master's degree or doctorate in philosophy (Schwitzgebel & Cushman, in prep). In an early part of the test, participants judged a pair of specific hypothetical moral dilemmas similar to the well-studied trolley problem (e.g., Fischer & Ravizza, 1992; Foot, 1967; Kamm, 1998; Thomson, 1985). In "push"-type cases, the agent had to apply direct physical force to a victim, using him as a tool to save five other people: for example, throwing a man in front of a runaway boxcar to stop it from hitting five people further down the tracks. In "switch"-type cases, the agent acted at a distance to save five people, with the side effect that one other person would die: for instance, switching a runaway boxcar away from the main track where five people were threatened, and onto a side-track where one person would die. The critical manipulation was to change the order in which these two cases were presented. Non-philosophers are more likely to judge switch-type harm to be as bad as push-type harm when viewed in the order "push/switch," but are more likely to judge switch-type harm to be less bad than push-type harm when the cases are viewed in the order "switch/push." It turns out that philosophers are just as susceptible as non-philosophers to this effect: The order in which they view the cases has a statistically significant and surprisingly large effect on their patterns of judgment.

The most important evidence came, however, at the very end of the test. We asked philosophers whether or not they endorsed the "Doctrine of Double-Effect" (DDE), a well-known principle in philosophy that draws a moral distinction between push-type and switch-type cases. The results were striking: Philosophers were about 30% more likely to endorse the DDE — an abstract, explicit moral principle — when they had previously viewed specific moral dilemmas in the order "switch/push" rather than "push/switch." This effect was just as strong among the subset of philosophers who specialized in ethics and had received a PhD.

These results have two important implications. First, they provide evidence that philosophers' endorsements of moral principles can depend substantially on their prior judgments regarding particular cases. Second, they demonstrate that philosophical training does not inoculate against the influence of morally 'irrelevant' factors (such as the order in which two cases are presented) on principled reasoning.[2] Reflective equilibration surely plays a critical role in the construction of explicit moral theories.

Nevertheless, it appears that intuitive processes of moral judgment influence philosophical theories in ways that are both powerful and unseen.

Complexity

In 1997, the United States Supreme Court announced a landmark decision upholding New York's ban on physician-assisted suicide.[3] The case turned on the merits of a simple comparative question: Is killing a person the same as allowing him or her to die? According to law, a physician must respect a patient's wish to withhold lifesaving medication — that is, doctors can be required to allow a patient to die. Defenders of a right to physician assisted suicide asserted that the distinction between active euthanasia (e.g., administering a lethal dose of morphine) and passive euthanasia (e.g., withholding a lifesaving dose of antibiotics) is nothing more than a semantic sleight of hand. Either way, they argued, the patient's death depends on the doctor's choice. But the Court disagreed. The majority opinion in Vacco v. Quill held that there is a significant moral distinction between actively killing and passively allowing a person to die.

The moral distinction between active and passive harm is well represented in the philosophical literature (Fischer & Ravizza, 1992) and has a large influence on ordinary people's moral judgments (Baron & Ritov, 2004; Cushman et al., 2006; Ritov & Baron, 1999; Spranca, Minsk, & Baron, 1991). From a certain perspective, however, it is hard to explain or to justify. Consider again Vacco v. Quill: In both cases the doctor's decision is unequivocally responsible for the patient's death. In both cases the doctor is doing what the patient wants. Why does it matter whether the death was caused by performing a physical act or, instead, by failing to act?

We use the distinction between actions and omissions as a case study of complexity in moral principles. We argue that actions typically support more robust, automatic attributions of causation and intention. Because these attributions constitute basic inputs to the process of moral judgment, the action/omission attribution affects moral judgment. In essence, the moral distinction derives from distinctions made by non-moral cognitive processes such as causal attribution and intentional attribution. More broadly, we suggest that much of the complexity of our moral rules ultimately derives from the complexity of non-moral cognition.

For example, consider John who rolls a ball toward 12 pins (an action) and Jane who stands by and allows the ball to roll (an omission). John might be considered more causally responsible for the pins' falling than Jane is, and also to have intended the pins to fall more than Jane did. This is an example of the action vs. omission distinction operating in non-moral attributions of causation and intention. Possibly, the action/omission distinction carries through to affect moral judgments in the context of harmful behavior because causal responsibility for harm and intent to cause harm are key determinants of moral judgments. Replace the 12 pins with an innocent child, and John might look morally more culpable than Jane because, in some intuitive sense, he appears to have caused the child harm and intended the harm more than Jane.

Experimental evidence supports this hypothesis. People's judgments of non-moral actions and omissions (e.g., the bowling case) do reveal systematic discrepancies in causal attribution and intentional attribution (Cushman, Young, & Hauser, in prep). Specifically, people assign more causal responsibility to actions than to omissions, and they are more likely to consider actions intentional. Indeed, there is some evidence that actions support more robust causal inferences about an agent's goal (similar to intent) even during infancy (Cushman, Fieman, Schnell, Costa, & Carey, in prep). In the relevant study, six- to seven-month-old infants watched as a hand repeatedly reached for and grasped a series of objects. In the "consistent action" condition, the hand always reached for one object (e.g., a ball), preferring it to any other object (a banana, a box, a watch, etc.). Infants in the action condition expected the hand to continue reaching for the ball, as revealed by the duration of their gaze to expected versus unexpected events. In the "consistent omission" condition, the hand never reached for the ball, preferring to reach instead for any other object. Infants in the omission condition failed to form any expectation about the hand's future behavior; they were entirely unsurprised to see the hand change course and prefer the ball to future objects. Thus, infants appear to infer goals from consistent actions ("he always goes for the ball") but not consistent omissions ("he never goes for the ball"), even when the evidence in favor of each inference is equal.

It is well known that moral judgments depend substantially, although not exclusively, on assessments of causal responsibility for harm and intent to harm (Alicke, 1992; Cushman, 2008; Darley & Shultz, 1990; Piaget, 1965/1932; Royzman & Baron, 2002; Young, Cushman, Hauser, & Saxe, 2007; see also Pizarro & Tannenbaum, this volume). Thus, harmful actions may seem morally worse than harmful omissions because the active agent appears to have caused and intended the harm more. To establish this causal connection between non-moral attributions and moral judgments, Cushman and colleagues (in prep) took advantage of the finding that the judgment of deserved punishment relies significantly more on causal attributions than does the judgment of moral wrongness (Cushman, 2008). If causal attribution is partially responsible for the moral distinction between actions and omissions,

then the action/omission distinction should exert greater influence on punishment judgments than on wrongness judgments. This is precisely what the study revealed.

Further evidence for the role of causal attribution in the action/omission distinction comes from a series of studies by Baron and colleagues (Asch et al., 1994; Baron & Ritov, 2004; Ritov & Baron, 1999; Royzman & Baron, 2002; Spranca et al., 1991). They consistently found that (1) many people explicitly state that actions are more 'causal' than omissions; (2) people who make that assessment are much more likely to judge harmful actions to be morally worse than harmful omissions; and (3) people explain their moral distinction between actions and omissions by appealing to the underlying causal distinction.

This evidence from people's explicit justifications for the action/omission distinction raises a key question: To what extent does the moral distinction between active and passive harm depend on an explicit, principled appeal to causal and intentional concepts rather than on processing features of implicit, automatic attributions of causation and intent? A recent study addressed this question by using functional neuroimaging to infer the cognitive processes underlying the action/omission distinction (Cushman, Murray, Gordon-McKeon, Wharton, & Greene, in prep). Of particular interest was activity in the dorsolateral prefrontal cortex (DLPFC), a region associated with the explicit, controlled application of abstract rules to a problem (Bunge & Wallis, 2007). The DLPFC was found to be significantly more active when subjects judged harmful omissions, compared to when they judged harmful actions. Taken alone, this evidence is compatible with either of the hypotheses we considered above. Possibly, activation in the DLPFC reflects the application of an explicit principled rule exonerating omissions: "The doctor is not responsible because he simply allowed the patient to die, but he didn't really cause the death." An alternative possibility is that activation in the DLPFC reflects the need to deploy controlled cognitive processes to condemn omissions: "The doctor is responsible because the patient's death depended on his purposeful decision." On this latter hypothesis, actions require less controlled, less deliberate DLPFC processing than omissions because automatic psychological mechanisms robustly condemn actions but not omissions.

These two hypotheses make opposite predictions about which participants will show the greatest amount of DLPFC activity when judging omissions. If the activity reflects the application of an explicit moral rule exonerating omissions, then people who show the greatest difference in judgment between actions and omissions should also show the most DLPFC activity. Alternatively, if the activity reflects the necessity of controlled processes to interpret and condemn harmful omissions, then people who show the smallest difference in judgment between actions and omissions should show the most DLPFC activity. This second pattern is what we observed: DLPFC activity during the judgment of omissions was significantly correlated with their condemnation. Thus, while people are able to report an explicit rule that accounts for the action/omission distinction after the fact, this study failed to provide evidence for the deployment of such a rule during the process of judgment itself. Instead, the evidence suggested that additional controlled, cognitive processing is necessary to equate harmful omissions with harmful actions. The automatic processes that support the judgment of harmful actions appear to be insufficient for the condemnation of harmful omissions.

In summary, it appears that the moral distinction between actions and omissions depends in part on non-moral processes of causal and intentional attribution. People—perhaps even young infants—tend to form more robust causal and intentional attributions from actions than from omissions. Automatic moral judgments rely on attributions of causation and intent as key inputs. Consequently, the non-moral action/omission distinction leads harmful actions to be judged morally worse than harmful omissions. As we saw above, consistent patterns of moral judgment constitute an important basis for the abstraction of general moral principles. In this way, the basic cognitive processes that young infants use to understand actions and events may contribute importantly to the moral doctrines endorsed by the US Supreme Court.

If the general structure of this argument is correct — if explicit moral principles reflect the processing features of relatively automatic, non-moral processes of causal and intentional attribution — then we can begin to explain the pervasive complexity of explicit moral principles. It derives from the much more general complexity of the cognitive mechanisms we use to interpret actions and events. By analogy, Pinker (2007) has argued that many of the complex rules governing the grammaticality of verbs depends on general (i.e., non-linguistic) processing features of those very same cognitive systems. From this perspective, moral rules and grammatical rules are two windows onto the basic structure of human thought.

Peculiarity

If the moral distinction between active and passive harm is an explicit formalization of our moral intuitions, why can it seem so peculiar? Active harm really does feel worse than passive harm. Yet, it still seems strange to say that it is worse for a doctor to

fulfill a patient's end-of-life wishes by actively doing something than by deliberately withholding treatment. In this section, we argue that this two-mindedness reflects the operations of dissociable cognitive systems that operate with incommensurable conceptions of causation, intention, and (consequently) morality (Cushman & Young, 2009; Cushman et al., 2006; Greene, 2008; Greene, 2001; Pizarro & Bloom, 2003; Sloman, 1996; White, 1990). When we construct an explicit theory of morality we do our best to capture the bases of intuitive judgments. But, the concepts available to our explicit thought processes are fundamentally mismatched to the underlying bases of our intuitive, affective responses. Because of this mismatch, intuitive patterns of judgment look peculiar to the explicit reasoning system that constructs moral principles.

To illustrate this phenomenon, we turn from active versus passive euthanasia to another, perhaps less well-studied, feature of the law. The Anglo-American legal tradition employs two distinct concepts of causation: "factual causation" and "proximate causation." Factual causation has a simple, clear definition that sounds entirely reasonable: A person's behavior caused an event if the event would not have occurred in the absence of the behavior. How do you know if Frank's shot caused Mary's death? It did so if Mary would be alive but for Frank's pulling the trigger. This 'but for' criterion gives factual causation its other popular name in law: the sine qua non test.

The problem with factual causation as a legal concept is that it completely fails to capture our intuitive judgments about moral responsibility—or, for that matter, causation. To see why, consider the following version of Frank's shot and Mary's death. When Frank pulled the trigger, his bullet hit the bull's eye. Frank, not Bruce, won the shooting competition. Without the $10,000 prize, Bruce cancelled his trip home for Thanksgiving. Too bad, because Bruce's aunt Mary choked on a cranberry, and Bruce's CPR skills undoubtedly would have saved her. But for Frank pulling the trigger, Mary would still be alive… and yet we not only refuse to hold Frank morally responsible for Mary's death; we don't even want to say that Frank caused Mary's death.

Unlike factual causation, proximate causation is a hopelessly complicated concept that utterly resists definition. As its name implies, one of the critical factors is proximity, which can be understood temporally, spatially, or in terms of intervening 'events'. (Recall that British law long held a person guilty of homicide only if his or her behavior caused death within a year and a day.) But perhaps the most peculiar element of proximate causation—and arguably the most fundamental—is foreseeablity. Suppose that Anne's daughter complains of an ear infection and a stomach ache. Anne asks her husband, a chemist, what to do. He suggests Tylenol for the ears and Pepto-Bismol for the tummy. Now, suppose that these two medications react to produce a very toxic substance, as any chemist would know. But Anne's husband is tired and fails to think very hard about his advice. Anne administers the medicine to their daughter, and her reaction to the toxic compound is fatal. Here is the critical question: Who caused the daughter's death? If your intuitions point towards Anne's husband, then you'll be glad to hear that the law does, too. An agent is typically considered the proximate cause of a harmful outcome only if a reasonable person in that agent's shoes would have foreseen harm as a likely outcome of his behavior. Anne could not reasonably be expected to foresee harm, but her husband could have.

Proximate causation is sometimes maligned by legal scholars who disapprove of a causal concept that cannot be defined and depends on factors, such as foresight, that do not seem to have anything to do with causation at all. Yet the law requires proximate causation because it succeeds brilliantly where factual causation fails: Proximate causation captures our intuitive judgments of causal and moral responsibility. At some level, it does not just capture our intuitions—it is our intuitions. Although attempts have been made to characterize proximate causation, it does not exist as a defined doctrine; rather, it is a collection of legal precedents born in the nuanced peculiarities of individual cases and gerrymandered to suit jurists' needs.

As you might expect, psychological theories of ordinary people's intuitive causal judgments resemble the legal concept of proximate cause. For instance, consider the role of mental state information in assigning causal responsibility. Lombrozo (2007) has demonstrated that adults are more likely to assign causal responsibility for an event to an agent who brings it about intentionally rather than accidentally. More recently, Muentener (2009) demonstrated that intentional actions are more likely to support causal inferences in infants. These findings are not an exact fit to the legal doctrine; the psychological studies implicate an agent's intention as a key element of causation, while the legal concept of proximate cause depends on what a reasonable person in the defendant's situation would have foreseen. But in each case, mental state representations exert an unexpected influence over intuitive causal judgments.

We have taken this detour through legal concepts of causation because they seem to parallel psychological mechanisms of causal judgment present in ordinary people. As we have seen, proximate causation captures elements of our intuitive causal judgments. Just as importantly, however, factual causation captures a prominent explicit causal theory (White, 1990). Philosophers and psychologists often refer to the 'but for' test that defines factual causation as a

'counterfactual' theory of causation. There are other popular explicit theories of causation as well. For instance, 'mechanistic' or 'production' theories of causation trace causal histories by exclusively tracing the transfer of energy through matter.

Critically, it appears that our explicit theories of causation are incommensurable with the psychological mechanisms that produce intuitive causal judgments. To put the point metaphorically, the words in our explicit causal language simply cannot express the ideas employed by our intuitive mechanisms of causal judgments. For instance, neither counterfactual nor production theories of causation have any place for mental state concepts such as foresight or intent, yet mental state factors play a critical role in our intuitive causal judgments. If you try to create an explicit causal theory that captured our intuitive causal judgments using only the conceptual resources available within counterfactual and production theories, the result will be both complicated and insufficient. Alternatively, you could construct an explicit theory that draws on representations of others' mental states, but then you would no longer recognize it as a causal theory. By the lights of our explicit causal theories, foreseeability simply does not belong. This may sound familiar: When legal scholars try to define proximate causation explicitly, what they end up with is complicated, insufficient, and alarmingly un-causal.

The incommensurability of explicit theories and intuitive mechanisms of judgment plays a key role in explaining why complex moral principles are generalized from what may be simple moral rules — and why they can look so peculiar. Let us suppose that intuitive moral judgments of harmful actions are generated by an extremely simple computation: An agent acted wrongly if her actions intentionally caused harm. Additionally, let us suppose that the representational inputs into this computation are intuitive attributions of causation and intention. Now, assume that a person attempts to generalize an explicit moral theory over his or her pattern of intuitive judgments. As a first pass, the person constructs the following theory: "An agent acted wrongly if his or her actions intentionally caused harm." But the available explicit theories of causation and intention will produce counter-intuitive moral judgments whenever those theories are at variance with their intuitive counterparts. This unfortunate person is now left trying to build an explicit moral theory that captures intuitive moral judgments, but using explicit concepts that are incommensurable with those embodied implicitly by the psychological mechanisms that determine his or her intuitive moral judgments. What the person ends up with is complicated, insufficient, and sometimes alarmingly un-moral. In short, he or she ends up with moral principles that look peculiar, like the distinction between active and passive euthanasia. In the final section, we ask why those peculiar-looking moral principles are so persistent.

Persistence

If moral principles like the action/omission distinction and proximate causation are so peculiar by lights of our explicit concepts, why do they persist? To bring the question into focus it helps to contrast moral principles with scientific principles. One of the enduring metaphors of the cognitive revolution is the 'person as a scientist'. The idea is that people have explicit theories (also called 'folk theories') that describe, explain, and predict the world around them. Of course, when we construct explicit theories about the world we are forced to rely on representational input from lower-level, automatic systems. Dennett (1991) argues against the notion of a 'Cartesian Theater', a removed vantage point from which a person watches his or her own mental processes. But if we take the notion of the 'person as scientist' seriously, there is such a vantage point: The theater is occupied by a scientist who is using controlled psychological processes to interpret the representational output of automatic psychological systems as they are projected into consciousness.

Consider, for example, an ordinary person's understanding of the laws of physical motion. A large body of evidence suggests that people have an intuitive sense of physical motion that operates with dramatically different properties than Newtonian mechanics (Caramazza, McCloskey, & Green, 1981; McCloskey, 1983). Consequently, when asked to produce an explicit theory of physical motion, people tend to produce an "impetus theory" remarkably similar to pre-Newtonian scientific theories. Just as in the moral cases presented in this chapter, people's explicit theories of physics can reflect processing characteristics of automatic, intuitive psychological mechanisms.

Yet folk theories, like scientific theories, can be revised, rejected, and reconstructed, ultimately moving beyond the structure of any particular automatic mechanism. This kind of conceptual change occurs when people check the predictions of a theory (e.g., an impetus theory) against the actual, represented phenomenon in the world (e.g., the motion of billiard balls). Empirical evidence clearly indicates that Newtonian mechanics explains reality better than impetus theory. So, if people are able to revise their explicit theory of physics, moving beyond the input of automatic mechanisms of physical understanding, why doesn't morality work the same way? We propose an answer to this question that depends on a distinction between representational and affective processes.

Explicit theories of scientific domains such as physics, biology, and psychology are representational. That is, folk theories are mental structures that map onto structures in the world and are used to describe, explain, and predict those structures. This representational function means that there is a place—the real world—from which new structure can be derived. Consequently, folk theories can take on representational structure exceeding that of the input mechanisms simply by comparing the internal representation (a folk theory of physics, for instance) to external events (actual physical events).

By contrast, affective processes assign value to things and events in the world, providing us with a basis for choosing between possible courses of actions. Morality is, at its core, an affective system (Haidt, 2001, 2007; see also Graham & Haidt, this volume). The function of morality is not to provide an accurate working model of events in the world, like a theory of physics. Rather, its functional role is to guide our behavior, telling us which behaviors of our own to inhibit and which to perform, when to punish people and when to reward them, etc.

Consequently, an explicit theory of one's affective responses—a theory of what is pleasurable, what is beautiful, or what is moral—will not reflect the structure of the physical world; it will reflect the psychological structure of our evaluations of it. Consider a moral claim such as "Killing babies for fun is wrong." This moral claim makes no straightforward predictions about the world that can be tested by an experiment. To the extent that it makes a prediction, that prediction concerns the structure of our minds. To say, "Killing the baby is wrong" predicts that killing the baby will feel wrong. This is a fundamental difference between theories formed over affective content and theories formed over representational content.

So where does this leave the scientist in the theater? He or she constructs explicit moral principles on the basis of moral intuitions designed to motivate. This process of construction may proceed very similarly to the construction of folk theories, but there is at least one key difference. Our representational folk-scientific theories can be checked against data outside our heads: biological structure, physical structure, etc. By contrast, the content of moral principles can be checked only against data inside our heads: the motivational mechanisms we use to make moral judgments. Thus, it is worthwhile to differentiate between two occupants of the theater: (a) the familiar scientist and (b) a philosopher. Like scientists' theories, philosophers' theories are carefully tested and revised. But the scientists' questions are answerable by testing and revising theories against data gleaned from the external world. By contrast, the philosophers' questions are answerable by testing and revising theories against data gleaned from the mind.

Applying this perspective to the particular case of the action/omission distinction, we are at last in a position to explain its complexity, peculiarity, and persistence. Automatic mechanisms of causal and intentional attribution respond more robustly to actions than to omissions. Consequently, automatic mechanisms of moral judgment yield a greater affective response prohibiting actions, as compared to omissions. This introduces some level of complexity to our moral judgments, although our explicit representational theories of causation and intent may reject a bright-line distinction between actions and omissions. Thus, in light of our explicit theories, discrepant moral judgments of actions and omissions look peculiar. The explicit moral distinction is persistent, however, because moral judgment is an affective process. While the judgment that "active euthanasia feels wrong, but withholding life-saving treatment feels okay" reflects the output of representational processes (e.g., causal and intentional attributions), the judgment itself is not a representation. This makes the explicit moral theory distinguishing actions from omissions difficult to revise or reject.

Conclusion

The metaphor of the philosopher in the theater situates philosophy within the ordinary person's mind. On the one hand, it provides a valuable lesson about the psychology of ordinary people. Just as ordinary people act like scientists, constructing, testing, and revising theories about physics and biology, ordinary people also act like philosophers, constructing, testing, and revising theories about right and wrong. On the other hand, it provides a valuable lesson about the nature of moral philosophy. Just as theories in scientific domains will tend to reflect the structure of the world, there is reason to suppose that moral philosophies will tend to reflect the structure of the mind. And to the extent that the content of explicit moral theories depends on widely shared patterns of intuitive moral judgments, we can explain three salient properties of law, policy, and philosophy: complexity, peculiarity, and persistence.

We conclude on a more circumspect note, considering a very large body of moral principles that our model does not explain. The rules of evidence in American courts are a useful example. They are certainly very complex, sometimes peculiar, and yet persistent. Among the most peculiar are rules that prevent juries from considering information that is highly reliable and relevant to a case when, for instance, it was improperly collected by police authorities (the exclusionary rule) or when it depends on second-hand rather than direct testimony (the hearsay rule). Although we have not studied other people's intuitive

judgments of evidential rules, the exclusionary rule and hearsay rule certainly violate our own intuitions. For the sake of argument, suppose that is strongly counter-intuitive to prevent a jury from hearing all the relevant evidence against an accused murderer, as we strongly suspect is the case. How can we explain the rules of evidence?

Our answer is neither surprising nor unique: Complex, peculiar rules of evidence persist because they work. We want to see as much relevant evidence as possible presented in trial; at the same time, we want to see safeguards against unscrupulous police practices or unreliable statements on the witness stand. Our rules of evidence strike a balance between these competing interests. We endure the counterintuitive outcomes of those rules in specific cases because they support a broader system that functions well, meeting our standards for the processes and outcomes of the law. There is a broader issue at stake in this example. We, and many others, have written about the ways that people construct and use moral principles that move beyond the raw material of their intuitive judgments of particular cases (e.g., Cushman & Young, 2009; Greene, 2008; Kohlberg, 1969; Pizarro & Bloom, 2003; Rawls, 1971, to cite only a few). For the purposes of this essay, a simple point suffices: Having situated the philosopher in the theater, we are in a better position to plot his or her escape.

References

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*, 368-378.

Asch, D. A., Baron, J., Hershey, J. C., Kunreuther, H., Meszaros, J., Ritov, I., et al. (1994). Omission bias and pertussis vaccination. *Medical Decision Making, 14*, 118-123.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*, 74-85.

Bartels, D. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition, 108*, 381-417.

Bunge, S. A., & Wallis, J. D. (Eds.). (2007). *Neuroscience of rule-guided behavior*. New York: Oxford University Press.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about the trajectories of objects. *Cognition, 9*, 117-123.

Cushman, F. A. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353-380.

Cushman, F. A. (2008). *The origins of moral principles*. Unpublished Doctoral Dissertation, Harvard University, Cambridge, MA.

Cushman, F. A., Fieman, R., Schnell, J., Costa, J., & Carey, S. (in preparation). Untitled manuscript.

Cushman, F. A., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (in preparation). Untitled manuscript.

Cushman, F. A., & Young, L. (2009). The psychology of dilemmas and the philosophy of morality. *Ethical Theory and Moral Practice, 12*, 9-24.

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

Cushman, F. A., Young, L., & Hauser, M. D. (in preparation). *Patterns of moral judgment derive from non-moral psychological representations*.

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

Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little Brown.

Fischer, J. M., & Ravizza, M. (1992). *Ethics: Problems and principles*. New York: Holt, Rinehart & Winston.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5-15.

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 3, pp. ???-???). 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*, 364-371.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144-1154.

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.

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

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology, 31*, 191-221.

Haidt, J. (2007). The new synthesis in moral psychology. *Science,* 316, 998-1002.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgment and justification. *Mind and Language, 22*, 1-21.

Kamm, F. M. (1998). *Morality, mortality: Death and whom to save from it*. New York: Oxford University Press.

Kamm, F. M. (2006). *Intricate ethics: Rights, responsibilities, and permissible harm*. New York: Oxford University Press.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 151-235). New York: Academic Press.

Kohlberg, L. (1981). *Essays on moral development, Volume 1: The philosophy of moral development*. New York: Harper Row.

Lombrozo, T. (2007). *Mechanisms and functions: Empirical evidence for distinct modes of understanding*. Paper presented at the 33rd annual conference of the Society for Philosophy and Psychology, Toronto.

McCloskey, M. (1983). Intuitive physics. *Scientific American, 248*, 122-130.

Mikhail, J. M. (2000). *Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A theory of justice'*. Unpublished Doctoral Dissertation, Cornell University, Ithaca, NY.

Muentener, P. (2009). *The origin and development of causal reasoning*. Doctoral Dissertation, Harvard University.

Paxton, J. M., & Greene, J.D. (in press). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*

Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental Science, 13*, 265-270.

Petrinovich, L., O'Neill, P., & Jorgensen, M. J. (1993). An empirical study of moral intuitions: towards an evolutionary ethics. *Journal of Personality and Social Psychology, 64*, 467-478.

Piaget, J. (1965/1932). *The moral judgment of the child*. New York: Free Press.

Pinker, S. (2007). *The stuff of thought*. New York: Viking.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: Comment on Haidt (2001). *Psychological Review, 110*, 193-196; discussion 197-198.

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

Ritov, I. I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes, 79*, 79-94.

Royzman, E., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research, 15*, 165-184.

Schwitzgebel, E., & Cushman, F. A. (in preparation). Untitled manuscript.

Shweder, D., & Haidt, J. (1993). The future of moral psychology: Truth, intuition, and the pluralist way. *Psychological Science, 4*, 360-365.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76-105.

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

Uhlmann, E., Pizarro, D., Tannenbaum, D., & Ditto, P. (2009). The motivated use of moral principles. *Judgment and Decision Making, 4*, 476-491.

White, P. A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin, 108*, 3-18.

Young, L., Cushman, F. A., Hauser, M. D., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*, 8235-8240.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: Guilford Press.

Footnotes

1. This essay draws largely from ideas and material in Cushman's doctoral thesis (Cushman, 2008). These have been revised and extended in the present essay with assistance by Greene.
2. At least, we presume that most philosophers would deem order of presentation irrelevant.
3. Vacco v. Quill, 521 U.S. 793.