The moral principles found in philosophy and embodied in law are often strikingly complex, seemingly 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 “causing another person to die.” In fact, the modern Model Penal Code requires pages of fine print to explain the details, just as it has over many decades and several comprehensive revisions.

Our goal in this chapter is to explain three properties of moral principles: complexity, peculiarity, and persistence. Our argument rests on a simple model of the origins of moral principles. Abstract, general moral principles are constructed from the raw material of intuitive judgments of particular cases, as explained in the next section of the chapter. Those intuitive judgments 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 ...”), which are pre-requisite to moral judgment (“… which was wrong”). Consequently, explicit moral principles reflect the complexity of our psychological processes of causal and intentional attribution, which we will explain in detail. That complexity can seem peculiar because our

---

1 This essay draws largely from ideas and unpublished material in Cushman’s doctoral thesis (Cushman, 2008). These have been revised and extended in the present essay with assistance by Greene.
intuitive, automatic attributions of causation and intent are often at odds with more considered, rational theories of causation and intent. But, the complex principles are persistent because they come with the strong emotional force of our intuitive “gut sense” of right and wrong. Emotions are uniquely difficult to revise or reject, an issue considered in the final major section of this chapter.

*Intuitions and Principled Reasoning*

We begin with the simple observation that ordinary peoples’ moral judgments often track philosophers’ and lawyers’ explicit principles, mirroring their complexity and nuance. For a long while, psychologists assumed that explanation of this phenomenon was simple: 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). At least in some cases, people make moral judgments that are consistent with prominent philosophical or legal principles, and yet they may 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 the pre-school years (Pellizzoni, Siegal, & Surian, 2010).

These recent findings invite quite a different model of the relationship between philosophers’ principles and ordinary people’s judgments. It seems that some philosophical principles are abstracted from the kinds of intuitive judgments readily observable in the daily lives of ordinary people, even of very young children. To some extent, this should not come as a surprise. Many philosophers commonly test their principles against intuitions formed in thought-experiments, and some openly embrace the project of systematizing moral intuition (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 strong, automatic, affective responses to particular cases (Haidt, 2001; Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009). In short, 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, 2008; Mikhail, 2000; Shweder & Haidt, 1993).

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 when the sister is infertile. They claim it is wrong because the family would be embarrassed; but they insist that it is wrong even when kept secret. They claim it is wrong because the siblings will regret it; but they insist it is wrong even when 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 about incest. It appears that their goal is not to develop a rational theory to guide future judgments, but instead to paint a veneer of reason over a foregone conclusion.

Yet there is evidence that people’s moral justifications are more considered and constrained than these early studies suggested. To begin with, it is worthwhile to note that typical justifications for the prohibition against incest focus on the possibility of harm (e.g., to the child, the family, or the siblings themselves). By now, a considerable amount of evidence
suggests that people use an explicit rule that, all else being equal, harm is bad and should be minimized (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) (Cushman et al., 2006; Kohlberg, 1981; Piaget, 1965/1932). Meanwhile, there are other factors that play an important role in intuitive judgments but that people are relatively unwilling to endorse. For instance, people’s intuitive moral judgments reliably indicate an important role for direct physical contact (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 physical contact in their intuitive judgments, they often reject this factor as a valid principled basis for moral judgment (Cushman et al., 2006). For instance, one participant wrote, “I guess I was fooled by the line-pulling seeming more passive than the man-pushing, but that view is hard to justify now.”

These studies indicate that moral intuitions play an important role in the development of explicit moral theories, but also suggest that people are picky about the specific moral principles they will endorse. In this regard, people seem to engage in what philosophers call “reflective equilibrium”: a tug-of-war between intuitive attitudes and rational 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 reflective equilibrium is surely welcome news to philosophers. While ordinary people’s attempts to construct moral principles can sometimes look haphazard or unjustified, perhaps philosophical training inoculates against ‘errant’ rationalizations by developing expertise in careful reflective equilibrium.
A recent study put this proposition to the test by comparing the relationship between intuitive judgments and moral principles in a large population 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.\(^2\) 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 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. Philosophers were asked whether or not they endorsed the “Doctrine of Double-Effect” (DDE), a well-known principle in philosophy that dictates a moral distinction between push-type and switch-type

\(^2\) The trolley problem has been extensively investigated in moral philosophy (e.g., Fischer & Ravizza, 1992; Foot, 1967; Kamm, 1998; Thomson, 1985) and moral psychology (e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2000; Petrinovich, O’Neill, & Jorgensen, 1993). In Schwitzgebel and Cushman’s study, participants viewed either the classic trolley problem or one of three structurally similar cases involving other life-and-death situations (e.g. drowning or burning).
cases. The results were striking: Philosophers were 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 philosophers who specialized in ethics and had a completed PhD as it was among those with other specialty areas and a Master’s degree. And it was replicated using a second, independent moral principle known as moral luck.

These results have two important implications. First, they provide evidence that philosophers’ endorsement of particular moral principles can depend substantially on their prior intuitive 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.3 Reflective equilibrium surely plays a critical role in constraining explicit moral theories. Nevertheless, it appears that intuitive processes of moral judgment can influence philosophical theories in ways that are both powerful and unseen. In the next section, we explain how these processes cause explicit moral principles to reflect processing features of non-moral cognitive processes, thereby causing the principles to inherit the cognitive processes’ complexity.

*Complexity*

In 1997, the United States Supreme Court announced a landmark decision upholding New York’s ban on physician-assisted suicide.4 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

---

3 At least, we presume that most philosophers would deem order of presentation irrelevant.
4 Vacco v. Quill, 521 U.S. 793.
suicide asserted that the distinction between active physician assisted suicide (e.g., administering a lethal dose of morphine) and passive physician assisted suicide (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 to die.

The moral distinction between active and passive harm is well-represented in the philosophical literature (Fischer & Ravizza, 1992), and it has been shown to play a large and consistent role in structuring ordinary people’s moral judgments (Baron & Ritov, 2004; Cushman et al., 2006; Ritov & Baron, 1999; Royzman & Baron, 2002; Spranca, Minsk, & Baron, 1991). From a certain perspective, however, it is hard to explain or to justify. Consider again the defendant’s claim in Vacco v. Quill: In both cases the doctor’s decision is unequivocally responsible for the patient’s death. Why do people judge active harm to be morally worse than passive harm?

Here, we explore a simple hypothesis: Actions typically support more robust, automatic attributions of causation and intention. Because these attributions constitute basic inputs into the process of moral judgment, the action/omission attribution has an impact on moral judgment. In essence, the moral distinction depends on processing features of non-moral cognitive processes such as causal attribution and intentional attribution.

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 he appears to have caused the child harm and intended the harm more than Jane.

Experimental evidence bears out this hypothesis. People’s judgments of nonmoral 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 showed a robust expectation that the hand would continue to reach for the ball, as revealed by the length of their looks to expected versus unexpected events. In the “consistent omission” condition, the hand always omitted to reach 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. This suggests that infants have an easier time inferring goals from consistent actions (“he always goes for the ball”) than from 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). If actions support more robust attributions of causal responsibility and intent, this may account for the moral distinction between active harm and passive harm. To establish this causal connection between non-moral attributions and moral judgments, Cushman and colleagues (in prep) took advantage of a recent finding that the judgment of deserved punishment relies significantly more on causal attributions than does the judgment of moral wrongness (Cushman, 2008). Thus, if causal attribution is partially responsible for the moral distinction between actions and omissions, then the moral distinction should be larger in the judgment of deserved punishments than in the judgments of moral wrongness. 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 omission by appealing to the underlying causal distinction. Convergent evidence from Cushman and colleagues (2006) shows that most people are able to provide justifications for the moral distinction between active and passive harm.

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 extra 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 at least 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 processes of moral judgment rely on these 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 in the previous section, 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 some of the pervasive complexity of those explicit moral principles. Moral complexity may be inherited 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 lenses through which we can perceive the structure of human thought. In the next section, we take up another question: Why is it that some of the moral principles projected through this lens look so peculiar?

**Peculiarity**

There is a certain paradox in our “intuitive” reaction to the moral distinction between active and passive harm that lurked in the background of the previous section. On the one hand, we argued that basic, early-emerging and relatively automatic mechanisms of causal and intentional attribution respond much more robustly to actions than to omissions. On the other hand, we argued that from a certain perspective the action/omission distinction looks “peculiar” because a doctor who deliberately withholds lifesaving treatment both causes and intends the patient’s death. So which is it? Is our brain wired to see a difference between actions and omissions or to treat them identically?

In this section, we argue that the answer is both. We are quite literally “of two minds” when it comes to concepts of causation, intent, and morality (Bargh, 1999; Cushman & Young, 2009; Cushman et al., 2006; Greene, 2008; Greene & Cohen, 2004; Pizarro & Bloom, 2003; Sloman, 1996; White, 1990). Automatic, intuitive systems of action and event understanding show differential sensitivity to actions and omissions; meanwhile, controlled, rational systems of action and event understanding show identical sensitivity. This is why the DLPFC showed heightened activity during judgments of harmful omissions among people who judged omissions to be as bad as actions: Those people were required to use controlled cognitive processes to arrive at the guilty verdict. Consequently, the action/omission distinction looks peculiar to one
of our minds: speaking very roughly, to the DLPFC. The ‘sense’ of the distinction is lost in the translation between automaticity and reason.

In order for us to develop this argument in more detail, it will help to take a broader perspective on the distinct features of automatic versus controlled processes of causal attribution. Here again, a useful starting point is 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, let’s return to Frank and Mary. 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 great 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, not only don’t we want 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.)

Perhaps the most peculiar element of proximate causation—but arguably the most
fundamental—is foreseeability. A pair of examples will help to illustrate its role. Suppose that
Anne’s daughter complains of an ear infection and a stomach ache. Anne asks her husband, a
chemist, what to do. Her husband suggests Tylenol for the ears and Pepto-Bismol for the
tummy. Now, suppose that any reasonably well-trained chemist ought to realize that these two
medications will react to produce a very toxic substance. But Anne’s husband is tired and fails
to think very hard about his advice. Anne and administers the medicine to their daughter, who
ends up in the hospital with a severe reaction to the toxic compound. Here is the critical
question: Who played a greater causal role in bringing about the toxic reaction? 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 foreseeability, which 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.
It should not be surprising, therefore, that psychological theories of ordinary people’s intuitive causal judgments resemble the legal concept of proximate cause in certain respects. 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 generating our intuitive causal judgments. If we tried to create an explicit causal theory that captured our intuitive causal judgments, using only the conceptual resources available within counterfactual and production theories, the resulting theory would probably be both complicated and insufficient. Alternatively, we could construct an explicit theory that draws on representations of others’ mental states, but then we would no longer recognize it as a causal theory. By the lights of our explicit causal theories, foreseeability simply does not belong. If this sounds familiar, there is a good reason: 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 conceptual resources that are incommensurable with 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**

Our model of the origins of moral principles resembles a train wreck in slow motion. Philosophers, legal scholars, and ordinary people construct moral principles based in part on their automatic, intuitive moral judgments. Those judgments reflect processing features of relatively automatic, intuitive processes of action and event understanding, such as causal and intentional attribution. Consequently, features such as the distinction between active and passive harm end up in our explicit moral principles. But some of the basic processing features of automatic action and event understanding seem to be incommensurable with our explicit, rational theories of causation and intent. Thus, distinctions like the action/omission distinction seem peculiar even while they capture some intuitive sense of right and wrong. If those distinctions are so peculiar, why do they persist?

To bring the question into sharper focus, it will help 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) has mocked 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 putting on a show.

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 is a better approximation of reality than impetus theory. These mechanisms of conceptual change explain why bad scientific theories do not persist.

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 an affective system. The function of morality is not to provide an accurate working model of events in the world, like a theory of physics. Rather, the functional role of moral judgment is to guide our behavior. It tells us when to punish people and when to reward them, or which behaviors of our own to inhibit and which to perform.

To be sure, the structure of human moral judgment has been shaped by features of the social world, guiding us towards adaptive patterns of behavior. Still, although morality is adapted to the social world, its function is not to represent the social world. Analogously, the feeling of happiness is certainly shaped by features of the world, guiding us towards adaptive habits of behavior; but the function of happiness is not to represent the world. Happiness does not represent whiskers on kittens, warm winter mittens, or any of our other favorite things. Rather, happiness motivates us to pursue companionship, warmth, and so on, because they contribute to adaptive patterns of behavior.

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 structure in the world, but rather structure in our minds. Consider a moral claim such as “Killing an innocent baby is absolutely wrong.” This moral claim does not make any prediction about the world that can be tested by an experiment—at least, not any experiment we know of. To the extent that it makes a prediction at
all, the prediction concerns how we will feel about smothering the baby. 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. For this reason, it may be worthwhile to differentiate between two occupants of the theater: the familiar old scientist, but also 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, by lights 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 passive euthanasia 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 philosophical inquiry. Just as theories in scientific domains will tend to reflect the structure of the world, there is reason to suppose that philosophical theories 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 standard rules of evidence in American courts provide 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 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. 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


