How We Read People's Moral Minds

Liane Young

Liane Young *received her BA in philosophy (2004) and her PhD in psychology (2008) from Harvard University, after which she did postdoctoral work in MIT's Brain & Cognitive Sciences Department. She is an assistant professor in the Department of Psychology at Boston College, where she studies the cognitive and neural basis of human moral judgment. Her current research focuses on the role of reasoning and emotions in moral judgment and behavior—employing the tools and methods of social psychology and cognitive neuroscience, including fMRI (functional magnetic resonance imaging), TMS (transcranial magnetic stimulation), and the study of populations of patients with cognitive and neural deficits.*

Every day, we observe and experience the effects of other people's actions. Often such observation and experience lead us to evaluate the actors themselves, especially in cases where their actions are harmful or helpful. Yet the harmful or helpful outcomes aren't all that matter. What matters more to us, in fact, is whether those outcomes were intended. We want to know what people were thinking when they acted. Did he know that selecting that option would erase my hard drive? Did she intend to dye my hair purple? Figuring out what's going on in someone's mind presents a challenge, to be sure, but a worthy one if we're to decide whether to forgive or condemn and how much—and who deserves our trust and friendship and who doesn't. How do we meet this challenge?

Investigating what goes on in the mind and brain when we make moral judgments can help unpack a simple rule we apply across multiple contexts, as participants in social relationships, as jurors in a courtroom, as parents or teachers of young children: It is morally wrong to intend harm. Recent work suggests that our moral judgment of another person depends on specific brain regions for reasoning about that other person's mental state (such as the person's belief or intent) as well as on other brain regions for generating emotional responses to such mental state content. Studies reveal robust individual differences among people in their judgments of forgiveness and blame—differences that correlate with differences in their brain activity. Other studies show impaired moral judgments in patients with neurodevelopmental disorders or brain damage, as well as changes in people's moral judgments when their brain activity is experimentally modulated.

Let's start with the role of a perpetrator's mental states in our judgment of the perpetrator. Consider the following scenario: Grace and her co-worker are taking a tour of a chemical factory. Grace stops to pour herself and her co-worker some coffee. Nearby, there's a container of sugar. The container, however, has been mislabeled "toxic," so Grace thinks that the powder inside is toxic. She spoons some into her co-worker's coffee and takes none for herself. Her co-worker drinks the coffee—and, of course, nothing bad happens. When experimental participants are presented with such a scenario, most say that what Grace did was seriously morally wrong and that Grace is seriously morally blameworthy, simply on the basis of her harmful intent.[1]

Participants also considered this alternative scenario. Near the coffee machine is a container of poison. The container has been mislabeled "sugar," so Grace thinks the powder inside is sugar. She spoons some into her co-worker's coffee. Her co-worker drinks it and falls down dead. Again, most participants judged Grace on the basis of her mental state. That is, they were prone to let Grace off the hook because of her false belief and her innocent intention.

In sum, we judge failed attempts to harm to be morally forbidden and accidental harm to be more-or-less permissible. We judge actions to be morally wrong when there is intent to harm, regardless of whether harm is done. When there is no intent to harm—when harm is done, but by accident—we tend to be lenient. These behavioral patterns reflect the importance of reasoning about people's beliefs and intentions for evaluating their actions in moral terms.

Moral luck

There is still, however, a significant difference in our moral judgments of accidents versus fully neutral acts (i.e., acts performed with neutral beliefs and intentions and resulting in neutral outcomes). People judge accidents as *somewhat* morally forbidden and assign *some* moral blame to the responsible agents. Grace is judged morally worse when she accidentally poisons her co-worker than in the fully neutral situation, in which the container marked "sugar" actually contains sugar—even though Grace's intent was the same in the two situations.

What differs appears to be a matter of luck: a lucky (good) outcome (e.g., sweetened coffee) versus an unlucky (bad) outcome (e.g., poisoned coffee). Many moral judgments show this "moral luck" asymmetry: Unlucky agents, who cause harm by accident, are usually seen as

[1] L. Young, et al., "The neural basis of the interaction between theory of mind and moral judgment," *Proc. Nat. Acad. Sci.* 104, 8235-40 (2007).

morally worse. Yet it also seems to most of us that moral matters ought not to be matters of luck. Is the case of accidental harm an exception to the rule that beliefs and intentions, not lucky or unlucky outcomes, determine an agent's moral status? What accounts for the different moral judgments we assign to lucky versus unlucky agents?

There are two candidate factors. For instance, unlucky Grace not only causes a bad outcome (poisoning her friend) but holds a false belief (that the powder is sugar). Lucky Grace not only doesn't cause a bad outcome but holds a true belief (that the powder is sugar). These dimensions (belief and outcome) have typically been confounded in investigations of moral luck, making it impossible to tell whether the asymmetry in moral judgments we recognize as moral luck is due to the difference between lucky (good) and unlucky (bad) outcomes or the difference between true and false beliefs.

To test for the relative contributions of beliefs and outcomes to moral judgments, my colleagues and I developed a new scenario, featuring "extra-lucky" agents—that is, agents who hold the same false beliefs as unlucky agents but, thanks to an extra stroke of luck, don't cause any harm. Extra-lucky Grace falsely believes that the powder is sugar when in fact it is poison. She spoons the poison into her co-worker's coffee. However, her co-worker puts the coffee down and forgets to drink it, so no harm occurs. We hypothesized that extra-lucky Grace here would still be judged morally blameworthy on the basis of her false belief, even in the absence of any harmful outcome. Indeed, just as we hypothesized, extra-lucky Grace was judged more like unlucky Grace (who held the same false belief) than lucky Grace (who caused the same neutral outcome). The difference between true and false beliefs mattered more for moral judgments than the difference between neutral and bad outcomes. Furthermore, we found that the moral judgments made by our participants were critically affected by their assessments of whether the agent was justified in holding a belief—for example, that the toxic powder was actually sugar.[2] False beliefs were judged to be somewhat unjustified; therefore, agents holding those beliefs were judged to be somewhat blameworthy. Experiments like this reveal that mental-state factors (e.g., the truth and justification of an agent's beliefs) matter when we're making moral judgments, even when we're blaming people for accidents.

---

[2] L. Young, S. Nichols, & R. Saxe, "Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know," *Rev. Phil. & Psychol.* (in press).

The neural basis of representing mental states for morality

Recent work has targeted the neural basis of our ability to reason about people's mental states when we make moral judgments. This work builds on earlier work identifying specific brain regions for reasoning about mental states in non-moral contexts—for instance, when we must predict or interpret other people's behavior. Much of this earlier work uses a task from developmental psychology—the false-belief task, designed to test mental state reasoning in young children. In one scenario, children watch as a character named Sally places a ball in a basket and exits the room. Another character named Anne enters and moves Sally's ball to a box. Sally then returns, and the children are asked where Sally thinks her ball is. Children under four generally go for the box, because they cannot represent Sally's mental state as distinct from the real state of the world. Older children and adults pass the test by doing just that—representing Sally's false belief.

The false-belief task has been used to identify brain regions that support mental state reasoning. A number of fMRI studies show that a group of brain regions is selectively recruited by the false-belief task, compared with those brain regions recruited for a "control" task, such as reasoning about where to locate objects in outdated photographs or maps, or other non-mental representations. Research by Rebecca Saxe and colleagues suggests that one of these brain regions—the right temporoparietal junction (RTPJ), a patch of cortex above and behind the right ear—is selectively active in processing information about people's mental states as opposed to other kinds of information.

How does the RTPJ help us reason about mental states when we make moral judgments? In recent work, my colleagues and I conducted brain scans of participants while they read moral scenarios, such as the ones featuring Grace and her coffee mishaps. In one experiment, we varied the order in which we presented Grace's beliefs and facts about her action's outcome. For example, in half the trials, participants would read first that the powder was sugar and then that Grace believed the powder was poison. We reversed this order for the other trials. What we found was that the magnitude of the response in the RTPJ depended on whether belief or outcome information was being presented; the neural response was selectively higher for information about beliefs than for information about outcomes.[3] This result suggests that the

[3] L. Young (2007); L. Young & R. Saxe, "The neural basis of belief encoding and integration in moral judgment," *NeuroImage* 40, 1912-20 (2008).

neural response is sensitive to the presence of explicit mental states like beliefs, and that the RTPJ in particular supports the encoding of beliefs that are relevant for moral judgment.

Is there a relationship between this neural response and the actual moral judgment made? Does higher activity in the RTPJ predict greater reliance on mental states for making moral judgments? When we scanned the brains of undergraduate participants, we observed individual differences both in the moral judgments they made and in the magnitude of the neural responses when they made the judgments. Some participants judged accidental harm as very blameworthy, while other participants judged accidental harm as not very blameworthy at all. More important, these differences correlated with differences in the neural response. Participants with a low RTPJ response—and a presumably weaker representation of agents' false beliefs and innocent intentions—assigned more blame to agents causing accidental harm (like Grace, who accidentally poisoned her co-worker). Participants with a high RTPJ response blamed agents less for causing accidental harm.[4] This correlation suggests that individual differences in moral judgment (i.e., the assignment of blame or forgiveness for accidents) are due at least in part to individual differences in specialized neural circuitry for reasoning about other people's beliefs and intentions.

Of course, the conflict between mental-state and outcome factors may account for why we sometimes find it so hard to forgive. The brain data suggest that the strength of the mental state representation—how we reason about an agent's belief or intent—helps to determine whether we offer forgiveness even in the face of serious harm. Notably, the conflict between mental-state and outcome factors may be resolved quite differently in individuals with compromised mental state reasoning—as in autism spectrum disorders, including Asperger's syndrome. Our ongoing work suggests that individuals with Asperger's are more likely to judge harms caused accidentally due to false beliefs as morally wrong. In the absence of robust mental state representations, moral judgment appears to be based on outcome factors, such as the amount of harm that's done.

Changing moral minds

[4] L. Young & R. Saxe, "Innocent intentions: A correlation between forgiveness of accidental harm and neural activity," *Neuropsychologia* 47, 2065-72 (2009).

Given what we know about the brain basis of morality, can we alter moral judgment by altering activity in target brain regions? We did just this in a recent study: We changed people's moral judgments by producing temporary "virtual lesions" in their RTPJs, using a neurophysiological technique known as transcranial magnetic stimulation (TMS). TMS induces an electrical current in the brain, using a magnetic field to penetrate the scalp and skull. In this study, we used brain scans to identify the RTPJ in each participant and a nearby control region not implicated in mental state reasoning. We then conducted two separate experiments, in which we examined the effects of TMS on the RTPJ and on the control region of each participant. Experiment 1 consisted of two twenty-five-minute TMS sessions ten days apart, during which a participant received TMS to the RTPJ in one session and TMS to the control region in the other. Immediately after each session, participants read and responded to two dozen moral scenarios much like the one featuring Grace. (This task lasted under twelve minutes, and the post-stimulation effects of TMS, given our parameters, lasted from about half to twice the stimulation time.) In Experiment 2, we modified the protocol so that new participants received very short (500-millisecond) bursts of TMS while making the moral judgment for each scenario. This allowed us to investigate the effect of disrupting RTPJ activity precisely at the time of moral judgment, after mental state information had been presented and encoded.[5]

As we hypothesized, we found that TMS to the control region made no difference in either experiment. However, TMS to the RTPJ made a significant difference in both experiments: Moral judgments were based less on mental states and therefore more on outcomes. TMS to the RTPJ did not *reverse* moral judgments: Attempted harms (harmful intent, neutral outcome) were still judged morally worse than accidents (neutral intent, harmful outcome). Crucially, though, disrupting RTPJ activity led to more lenient judgments of failed attempts to harm, based on the neutral outcome, and harsher judgments of accidents, based on the harmful outcome.

Disrupting the neural processes that enable us to represent harmful intent, independent of harmful outcome, changes our moral judgments. Since moral judgments depend on specific neural substrates for processing information about beliefs and intentions, this aspect of morality can be selectively impaired by disrupting the specific neural processes for mental state reasoning.

---

[5] L. Young, et al., "Disruption of the right temporo-parietal junction with TMS reduces the role of beliefs in moral judgments," *Proc. Nat. Acad. Sci.* 107, 6753-8 (2010).

The role of emotions

The role of mental states in moral judgment appears to be one that we endorse; we regard the rule "It is morally wrong to intend harm" as rational. Is there a role for emotions in how we apply this rule? Recent fMRI and neuropsychological evidence suggests that moral judgments do depend on emotional responses to certain mental states. Moral judgments of attempted harms correlate with activation in the ventromedial prefrontal cortex (VMPC), a brain region for emotional processing, situated behind and between the eyes. The VMPC response correlates with the assignment of blame for harmful intentions in the absence of actual harm (e.g., as in a failed murder attempt). Individuals with a high VMPC response assigned more blame for failed attempts than did individuals with a low VMPC response.[6]

Does damage to the VMPC reduce moral blame for harmful intentions? We examined moral judgments made by patients with damage to the VMPC, patients with damage to brain regions not implicated in emotional processing, and healthy participants with no brain damage (all alike in age, gender, and IQ). VMPC patients judged attempted harm as significantly less wrong than the other participants did—and even less wrong than accidental harm. All nine of the VMPC patients we tested showed this same striking reversal of the normal pattern of moral judgments, judging failed attempts to harm as less wrong than accidents—revealing an extreme "no harm, no foul" mentality.[7] In line with previous work showing deficits in emotional processing of abstract versus concrete information, we found that VMPC patients may be unable to generate a normal emotional response to another person's mental state.

The ghost or the machine?

All of these studies focus on how the brain computes one important aspect of morality: the mental state of the moral agent. In general, moral neuroscience aims to uncover many different aspects of morality in the brain in their full physical glory. The work I've described here and the work of many others have begun to reveal that morality takes up space in the brain, and a lot of it. After all, morality depends on many cognitive functions—such as the ability to reason about people's intentions as well as about the outcomes of their actions, and to generate emotional

---

[6] Young & Saxe (2009).

[7] L. Young, et al., "Damage to prefrontal cortex impairs judgment of harmful intent," *Neuron*, 65, 845-51 (2010).

responses to that information. That the moral mind is rooted in the brain may strike us as scary. Is morality all machine and no ghost? Indeed, what we've learned from moral neuroscience so far suggests that how we behave and how we judge other people's behavior can be understood in neural terms—if not now, then eventually. Morality itself, though, may or may not be "all in our heads." Do moral truths, or moral facts of the matter, exist independent of how we think about them? Whether or not they do, the aim of science is to figure out how, in our minds and brains, we attempt to track them.