PSYCHOLOGY

The Declaration of Independence – A Lab Report

RESEARCH AT THE FRONTIERS OF PSYCHOLOGY IS EXPLORING WHY MORAL TRUTHS SEEM "SELF-EVIDENT"

By Fiery Cushman

The Declaration of Independence is an exceptional expression of moral conviction, but what makes it so? The document contains a list of the king's crimes and transgressions; a list that is lengthy, detailed, and now mostly forgotten. What we remember instead is a curious appeal to our natural sense of justice: "We hold these truths to be self-evident." Jefferson charges his king and country with having become "deaf to the voice of justice," a voice he believed to be native to mankind. His invocation of our sense of justice strikes a chord that still resonates today, more deeply perhaps than appeals to the authority of law, logic, history, or divinity. Sometimes justice is a matter of common sense – an intuition that defies justification or analysis, and seems to rise from the heart of our human nature.

Psychological research is beginning to catch up with Jefferson's rhetoric, uncovering the mental processes and brain structures that underlie our ethical intuitions. Many of our moral judgments emerge rapidly, forcefully, and apparently without conscious deliberation. Evidence about the structure of our "moral sense" is accumulating from diverse sources, from studies of adults, children, distant cultures, psychopaths, brain-damaged patients, and magnetic scans of the mind in action. But the picture that is emerging is a complex one, placing moral judgment at the nexus of numerous brain systems that interact and sometimes compete to produce our feelings of right and wrong. Jefferson was right to assume a basic, intuitive moral sense, but it is one shaped by learning, maturation, conscious reflection. and cultural influences.

To begin exploring what psychology has to offer philosophers, let's turn to a rather unphilosophical case study. Among the best candidates for a self-evident

James Turner

moral truth in the Jeffersonian sense is one that was surely far from his mind while penning the Declaration of Independence: our aversion to incest. Taboos against sexual intercourse between first-degree relatives are nearly universal among the world's cultures, and the historical record suggests this has long been true. In fact, many animal and even plant species have biological adaptations designed to avoid inbreeding. These mechanisms prevent organisms from investing costly resources in children whose genetic profiles would lack the necessary variation between paternal and maternal lineages that contributes to healthy development, silences genetic disorders, and wards off disease.

The biological disadvantages of incest are clear, but how is its avoidance instantiated in the human mind? Do we avoid intercourse with relatives after carefully reasoning about the dangers of resource allocation toward genetically unfit offspring, or by some more subtle influence? The social psychologist Jonathan Haidt has investigated this question by asking people directly about their views on incest. He tells his experimental subjects about a brother and sister who have intercourse once, enjoy it without regrets, keep it a secret, and use contraceptive protection. By and large, Haidt's subjects regard this behavior as exceptionally immoral. When Haidt asks them why they hold this view,

however, the subjects falter. They will attempt to identify a harmful feature of the behavior, only to discover that the scenario is carefully constructed as a victimless crime. Exasperated, many end up saying something like, "I don't know, I can't explain it. I just know it's wrong."

Haidt calls this phenomenon "moral dumbfounding" and suggests that it is a ubiquitous feature of human ethics. In his and others' experiments, subjects not only ultimately fail to provide justifications for firmly held moral beliefs, they go further, actually concocting bogus justifications in an attempt to rationalize their intuitions. Haidt suggests that a large part of what we accept as the deliberative foundations of moral judgment is in fact a scaffolding of spurious reasons hastily constructed around our intuitive moral sense.

The biological disadvantages of incest are clear, but how is its avoidance instantiated in the human mind?

The discovery of a substantial role for intuition in moral decision making might be regarded as unsurprising; after all, theories of unconscious psychological processes have been around as long as the field of psychology itself. In fact, however, the intuitionist perspective contradicts decades' worth of theories of moral psychology that emphasize the role of careful deliberation and conscious reasoning. All these theories trace their lineages back, in one manner or another, to the seminal work of Lawrence Kohlberg.

Kohlberg's doctoral dissertation, completed in 1958, outlined a pattern of moral development through which children pass with striking regularity. Maturation was marked by a progression through six stages of increasingly sophisticated moral reasoning, although not even mature adults were assured of reaching the highest level. Indeed, the sixth stage was so infrequently attained that it was sometimes omitted from later versions of the theory. Kohlberg, a Harvard professor, devoted his career to the exploration of the maturational process outlined in his dissertation.

Kohlberg established his stage theory by asking subjects to respond to moral dilemmas. In his famous Heinz dilemma, for instance, subjects are asked whether the impoverished Heinz is permitted to steal overpriced medications from the local pharmacy in order to save his dying wife. Whether the subject judged the behavior to be permissible or not was almost incidental to Kohlberg's research; what he wanted to understand was the process of reasoning that produced the judgment. In the "earliest" stages of development, subjects reason through the problem by applying simple and formulaic rules, such as "don't ever steal." Morality is viewed as a set of external constraints on behavior, like the commandments handed down to Moses. In the "middle" stages of development, subjects shift toward a social perspective on morality, as a set of rules agreed on by a community for mutual gain. Subjects might reason, for instance, that fair prices are essential to the relationship between buyers and sell-

ers, a principle which the pharmacist has violated. Finally, in the fifth and sixth stages of development, subjects conceive of morality as a requirement of rational behavior, paralleling Kant's conception of ethical behavior as the sole logical expression of free will. Subjects might reason that we would never choose rationally to construct a society in which laws governing theft supersede the obligation to protect one's family.

Kohlberg's theory can accommodate the idea of a "moral sense," but it is a sense profoundly different from the intuitivist conception offered by Haidt. Kohlberg's moral sense is a collection of abstract principles that unfold in a prescribed series of developmental changes. His moral sense is a guide to deliberate, effortful reasoning; subjects are fully aware of the moral principles that support their judgments. In this respect, Kohlberg's theory could hardly contrast more with the intuitivist picture of rapid, unconscious mechanisms giving rise to moral judgment.

How can we choose between these competing theories of our moral sense, or can they somehow be reconciled? At its core, the rift between the rationalist and intuitivist perspectives lies in their conceptions of moral justification. Rationalists view justifications as accurate reports of conscious reflective thought. Intuitivists view justifications as post hoc rationalizations of gut reactions. In order to understand the relationship between moral judgments and their subsequent justifications, it is necessary to develop experimental methods through which the two can be compared side by side.

A team of researchers at Harvard, comprising Marc Hauser, Liane Young, me and several others, set out to begin to answer these questions by borrowing from philosophical methods. The standard currency of moral philosophy is principles: clear rules that establish which behaviors are right or wrong. Our goal is to determine whether the moral principles discussed in the philosophical literature are accurate descriptions of the moral judgments of ordinary people, and if so, whether they operate by conscious reasoning or by intuition.

Philosophers often try to establish the validity of a proposed principle by developing a pair of test scenarios that differ in ways that appear to be morally significant, and arguing that one of the cases is *prima facie* more or less permissible than the other. Consider the following example. Case 1: Denise is on a trolley when the conductor goes unconscious. The trolley is heading toward five people on the main track where it will hit and kill them. Denise's only course of action is to flip a switch, sending the trolley down a side track where it will hit and kill one individual. Denise flips the switch. Case 2: Frank is standing by the trolley tracks when he witnesses a trolley running out of control toward five

people. Frank's only course of action to save the five is to push a fat man next to him onto the tracks, killing the man but slowing the trolley sufficiently to save the five. Frank pushes the man.

Philosophers have traditionally used this pair of scenarios to support the principle that it is less permissible to inflict harm as a means to accomplishing a goal (as in the case of Frank) than to inflict harm as the side-effect of accomplishing a goal (as in the case of Denise). They argue that Frank's behavior is worse because he intends the death of the one in order to save the five, while Denise merely foresees the death of the one as a side-effect of saving the five. Different philosophers have given this principle different names – for now, let's call this the "intention principle."

Cases like this one provide a unique opportunity to put Kohlberg's and Haidt's theories to an empirical test. By gathering subjects' responses to these cases, we can understand the principles at work in generating moral judgments. We can then compare these principles to the justifications that subjects provide for their moral judgments. If the principles that the subjects use to support their justifications align with the principles they use in their moral judgments – that is, the intention principle – then this counts in favor of Kohlberg's model according to which rational deliberation gives rise to moral judgments. If the principles that subjects provide in their justifications fail to align with the principles they use in their judgments, however, then it counts in favor of Haidt's model, in which our moral judgments arise intuitively from unconscious processes.

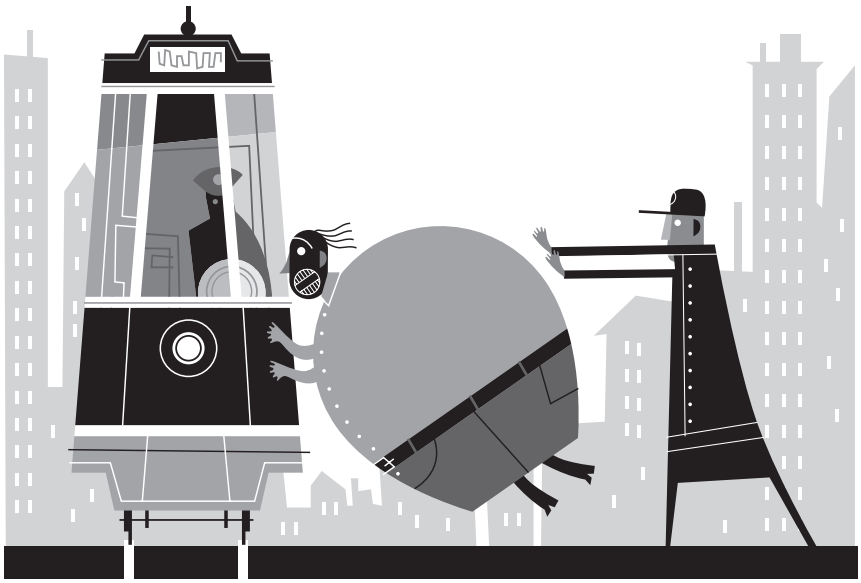
We put these cases on our research web site (www.moral.wjh.harvard.edu)

James Turner

and surveyed thousands of individuals from diverse cultural backgrounds. The results were unequivocal: nearly 90 percent of subjects judged Denise's action to be permissible while hardly more than 10 percent of subjects judged Frank's action to be permissible. The judgments of our subjects are clearly quite similar to those of philosophers, apparently conforming to the intention principle. Critically, however, only 30 percent of subjects were able to provide a sufficient justification for their patterns of judgment. Asked to distinguish between the two cases, many subjects said something like, "Frank is not permitted to kill one person, but Denise is forced to choose the greatest number of lives to save." This response fails to recognize, of course, that Denise also killed one person and that Frank also chose the greatest number of lives to save.

The inability of most subjects to come up with moral principles that adequately account for their judgments of these cases provides critical support for intuitivist theories. Further support came from our analysis of the cultural and demographic background of our test subjects. Subjects' judgments of these cases did not differ by age, level of general education, or exposure to moral coursework, as might be predicted if conscious reasoning were playing a strong role. Neither did judgments vary by nationality, religious background, ethnicity, or gender, although the sample was restricted to literate, English-speaking internet users.

On its face, our results seemed to endorse a rich version of Jefferson's self-evident truths: an unconscious mechanism for generating robust intuitions of right and wrong, operating in the absence of education and with apparent



James Turner

cross-cultural consistency. But something about this picture seemed incomplete. Among the self-evident moral dictates of Jefferson's age were human bondage, strict gender inequality, and the restriction of the vote to wealthy landholders. These values have lost the ring of truth, to say the very least – cultural variation is clearly an important feature of moral systems. Moreover, what do we make of Kohlberg's legacy? The rationalist picture of moral psychology may be incomplete, but the basic research findings refuse to go away. If you replicate Kohlberg's methods, you will replicate his results: people really do progress through predictable stages in their moral reasoning.

We went back to the drawing board, creating more than thirty new dilemmas that targeted not just one but several moral principles. These scenarios were designed to be more tightly controlled than the Frank and Denise experiment. Philosophers had tolerated certain imperfect features of the Frank and Denise cases, such as the fact that Frank pushes his victim with his own hands while Denise does her victim in with the pull of a lever. As it happens, such apparently minor differences have a big impact on moral judgments, and we took pains to eliminate them from our scenarios to ensure that subjects' judgments depended on the principles we wanted to target, and not on any other factors.

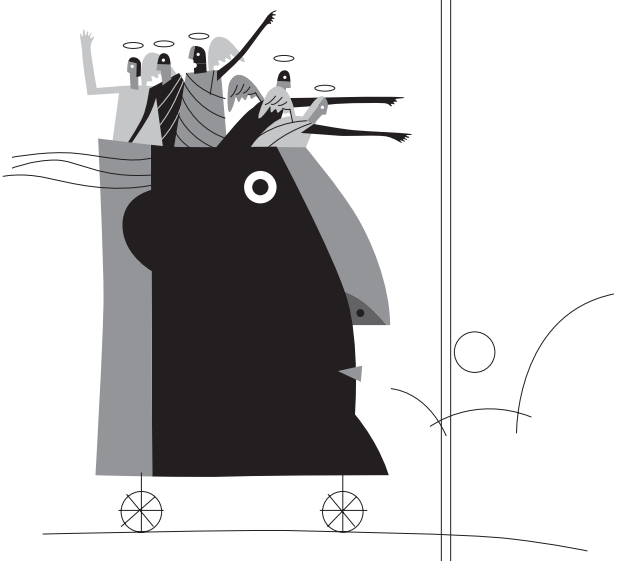
Human moral judgment is accomplished by multiple systems acting in concert.

When the results came in, they provided a more nuanced picture of the mechanisms of moral judgment. For the intention principle – the difference between harm as a means and harm as a side effect that was targeted in the cases of Frank and Denise – subjects still failed to provide sufficient justifications even for our more tightly controlled scenarios. But for other moral principles, subjects were perfectly able to provide sufficient justifications. For instance, subjects were able to articulate the difference between actively doing harm and passively allowing harm, a moral distinction that plays an important role in debates over euthanasia. Most people judge it morally worse to actively end a life by administering a lethal injection of toxic compounds, for instance, than to allow a life to end by not administering drugs to prevent the buildup of toxic compounds. For this principle, which we termed the "action principle," Kohlberg's model of rational deliberation seems a more successful explanation.

What is particularly striking about these results is that they demonstrate intuitive and conscious processes of moral judgment interacting side by side, often in one subject's appraisal of a single scenario. The lesson may not be surprising, but it is important: human moral judgment is accomplished by multiple systems acting in concert, some better characterized by intuitivist models and others better characterized by rationalist models. In the end, it's a lesson that both Kohlberg and Haidt would be comfortable with – neither was blindly

committed to a single perspective, and Haidt in particular has theorized about the interface of intuitive and rational processes.

Of course, it's one thing to know that multiple mechanisms interact to produce moral judgments, but it's quite another matter to understand how these mechanisms actually work. Often, an important first step toward a mechanistic understanding of mental processes is to target the brain regions that accomplish the task in question. A team of researchers at Dartmouth College used functional magnetic resonance imaging, or fMRI, to watch the brain in action as subjects made decisions about moral dilemmas. The moral dilemmas that the Dartmouth team chose targeted the same principles we used in our web study, the intention principle and the action principle. As could be predicted, these two moral principles activated different brain regions. In the case of the action principle – which subjects are able cite in their justifications and which therefore appears to involve conscious reasoning – the Dartmouth team found more activation in the dorsolateral prefrontal cortex, a brain region implicated in problem solving and abstract thought. In the case of the intention principle – which subjects are not able to cite in their justifications and which therefore appears to operate intuitively – the Dartmouth team found more activation in the orbitofrontal cortex and temporal pole, brain regions implicated in the experience of emotion.

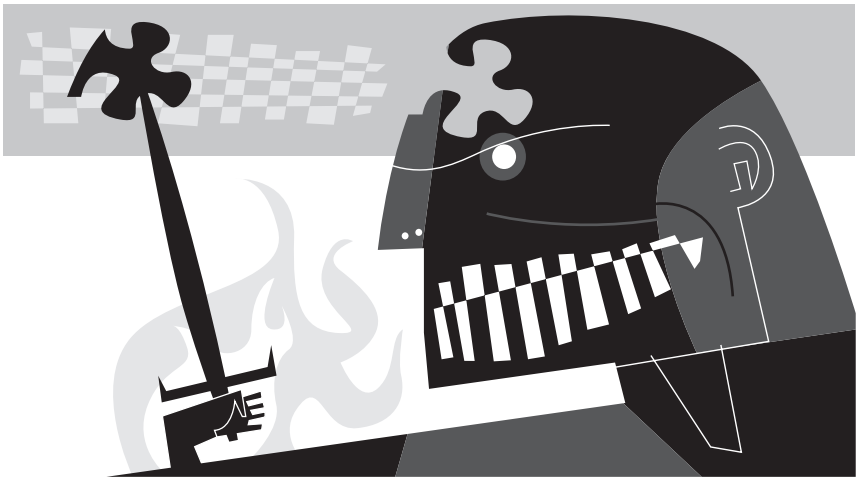


These results provide an important line of corroborative evidence that the intention and action principles are supported by dissociable brain systems. They also go further, offering tantalizing clues about the mechanisms underlying each system. Of particular interest is the association between the operation of intuitive moral principles and the activation of brain areas that process emotions. Is it possible that our emotions act as a moral compass, guiding us to intuitive feelings of right and wrong?

A version of this argument is being promoted by Joshua Greene. While studying toward his Ph.D. in moral philosophy at Princeton in the late 1990s, Greene became interested in the psychology of moral decision making. He teamed up with the cognitive scientist Jonathan Cohen to conduct one of the

James Turner

first fMRI studies of moral judgment, paving the way for the later work of the Dartmouth team. Greene was particularly interested in why people make moral choices that do not maximize their well-being. Some philosophers argue that maximal well-being must always dictate the proper moral choice. This branch of philosophy, today known as "consequentialism" because of its focus on the consequences of actions, is a variant of the utilitarian philosophy of Jeremy Bentham and John Stuart Mill. The chief rival to consequentialism is "deontology," a branch of philosophy, associated with Immanuel Kant, that emphasizes the inherent moral value of actions without exclusive regard to their consequences. A common strategy of deontologists is to present situations in



which the maximization of welfare seems to be clearly forbidden. One such classic situation is that of Frank, forced to choose between allowing five deaths or pushing a man in front of a train in order to prevent them. Deontologists argue that consequentialism fails in such cases because our clear intuition is that Frank must abstain from pushing the man, foregoing the welfare-maximizing choice.

In a series of experiments, Greene scanned images of people's brains while they made choices about cases like Frank and Denise. He discovered relatively higher levels of activation in emotion-processing areas of the brain when deontological choices were being made, and relatively higher levels of activation in general reasoning areas of the brain when consequentialist choices were being made. Greene's conclusion is that consequentialist decisions are driven by rational thought processes involving the maximization of welfare, while deontological decisions are driven by a forceful emotional aversion to directly causing the harm of specific individuals. Like our web experiments and like the

James Turner

Dartmouth group's fMRI study, Greene's data suggests that moral decision making involves interactions among several different brain systems, some associated with rational thought and others with gut feelings.

Greene's data also provides a potential insight into the rich body of evidence on individuals who have suffered trauma to areas of the brain that process emotions. The most famous such case is that of Phinneas Gage, a hapless worker on the Vermont railroad in the 1840s. In an extraordinary accident, Gage sparked a small explosion that blasted an iron rod through the front of his skull and thirty yards beyond. Even more extraordinary, Gage survived the injury without notable deficits in cognitive functions like motor control, language, and general abstract reasoning. What Gage did lose was a portion of his orbitofrontal cortex, a brain region implicated in the processing of emotions, among other functions. This brain damage led to a sudden and irreversible change in Gage's personality. He became temperamental and profane, and subject to violent outbursts that contemporaries described as animalistic.

Owing greatly to the work of Antonio Damasio and his collaborators at the University of Iowa, we are beginning to understand the condition of Gage and others with similar patterns of brain damage. The emotional deficits of these individuals seem to result in impairments in proper social functioning, including morally appropriate behavior. Some researchers have even gone so far as to describe the condition of individuals like Gage as a form of acquired psychopathy; the similarities between the symptoms of these brain-damaged individuals and the violent, manipulative, and antisocial behavior of psychopaths are striking indeed. An association between impaired emotional processing and socially inappropriate behavior is precisely what Greene's theory predicts: if certain emotions play a role in generating some of our key moral intuitions, individuals lacking those emotions ought to exhibit abnormal patterns of moral judgment, leading to aberrant behavior.

The similarities between the symptoms of these brain-damaged individuals and the violent, manipulative, and antisocial behavior of psychopaths are striking indeed.

But there is another possible interpretation of both Greene's neuroimaging results and the data on individuals with orbitofrontal damage. Rather than postulating a causal role for emotion in generating moral judgments, we might suppose that moral judgments in fact precede emotions, and that the role of emotions is to motivate behavior on the basis of these judgments. Such an interpretation makes sense of Greene's data, since we would expect the activation of emotional areas of the brain following moral judgments; and it makes sense of the patient data since, lacking the motivating force of emotions, these individuals would fail to make socially appropriate choices. In essence, this alternative proposes that we chart a course toward moral behavior with our moral sense

for a rudder and emotion providing the necessary wind. Greene's hypothesis goes a step further, actually allowing emotions to have a hand on the wheel.

In order to decide among these competing accounts, it is necessary to return to individuals with the appropriate profile of brain damage and test their moral judgments directly, rather than looking at the behaviors that may or may not be influenced by those judgments. To do this, our research group collaborated with several investigators specializing in individuals with frontal-lobe damage. The subjects we selected had damage in the ventromedial prefrontal cortex (vmPFC), and each had corresponding deficits in emotional processing. We asked them to judge a set of scenarios drawn from Greene's original fMRI study, cases similar to those of Frank and Denise. In cases where normal subjects typically make the consequentialist choice, favoring the welfare-maximizing outcome, the vmPFC group exhibited a perfectly normal pattern of responses. But in cases where normal subjects reject the consequentialist choice and instead favor deontological principles prohibiting direct harm, the vmPFC group diverged. Lacking an emotional aversion to pushing a man in front of a train, for instance, they were significantly more likely than normal subjects to consider this action an acceptable means of saving five individuals farther down the tracks.

When lightning strikes a man, we are certain to feel both sad and angry, but unlikely to experience moral outrage toward the cloud.

The results of this experiment lend strong support to Greene's hypothesis and suggest that emotions really do play a critical role in shaping our moral judgments – that they are in some sense constitutive of our moral sense. At the same time, it would be a mistake to suppose that our sense of what's wrong reduces to basic emotions such as sadness or anger. When lightning strikes a man, we are certain to feel both sad and angry, but unlikely to experience moral outrage toward the cloud. Emotional reactions must interface with cognitive systems that support our understanding of causation and responsibility, of other people's beliefs and desires, and of agency and free will – those systems that underlie our use of principles like the intention principle and the action principle. Here, too, the moral is that multiple systems are at work. Emotions contribute to our intuitive sense of morally impermissible behavior, but they must do so in concert with other cognitive systems. And, ultimately, these gut reactions compete against rational principles and learned moral rules, like the consequentialist, or utilitarian, commitment to maximizing welfare and happiness.

The more we learn about the human moral sense, the more we come to appreciate its complexity. Contrary to the popular fear that psychological research robs human behavior of its rich and individual character, the greatest benefit of this research is to provide a window onto the intricate and coordi-

nated activities of the mind that lie outside our everyday awareness, or that are obscured in the shadows of familiarity. We decide matters of right and wrong day in and day out, all our lives, but we are only just beginning to understand the full array of psychological mechanisms that contribute to this behavior. Morality is shaped by innate biases, cultural norms, and learned rules; it is the product of automatic intuitions as well as deductive reasoning; it depends upon our emotional responses, but acts together with our cognitive appraisals of causation, intention, and agency. Research into these areas holds a mirror up to the face of our humanity, rendering Jefferson's rough sketch in sharp detail.

But as we become more familiar with the image of our own moral sense, we will be forced to grapple with its implications. Should knowledge about the way we arrive at moral judgments shape the moral judgments that we make? Are we ready to fold a scientific understanding of morality into real-world systems of justice? These questions demand great care and consideration, to which, for the moment, I will contribute only a specific example. Recall the unique profile of consequentialist moral judgments exhibited by individuals with damage to the vmPFC. If these individuals perform well-intentioned actions that accord with their sense of justice but violate our own, how should society respond? Psychologists characterize such individuals with terms like "impaired" and "deficient," but these seemingly pejorative terms are only meant to capture the differences in their judgments, not to evaluate them as better or worse. In fact, the response of many philosophers to the vmPFC profile of judgment is to applaud it as clear, logical thinking untainted by the corrupting influence of emotion. It remains to be seen how these questions will be answered by policymakers, legislators, and the courts. As psychological facts about our sense of justice accumulate, the moral truths can seem less than self-evident.

FIERY CUSHMAN *is pursuing his doctorate in psychology in the Cognitive Evolution Laboratory at Harvard University. You can participate in online studies and learn more about the research project at www.moral.wjh.harvard.edu.*

James turner