Midwest Studies in Philosophy, XXXI (2007)

Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions

EDDY NAHMIAS, D. JUSTIN COATES, AND TREVOR KVARAN

> An agent would not be morally responsible at all if he was caused necessarily, predetermined, to try to do what he did, by his brain state . . .
>
> Richard Swinburne (1989, 51)

> But doesn't a truly scientific, mechanistic view of the nervous system make nonsense of the very idea of responsibility?
>
> Richard Dawkins (*Edge.org* 1/1/06)

> And since all behavior is caused by our brains, wouldn't this mean all behavior could potentially be excused?
>
> Jeffrey Rosen (*NY Times* 3/11/07)

> If you have no conscious control over what you do, how can you be held responsible for anything you have done? [i]t wasn't you, it was the chemical reaction that did it.
>
> Folk #95193 (5/1/07)

> [n]o [free will], the neural processes make the decisions.
>
> Folk #69709 (4/28/07)

As these quotations illustrate, it seems that people—from philosophers to scientists to journalists to the ordinary "folk" we have surveyed—share the intuition that "if our brain makes us do it, then we aren't morally responsible." We think that this

intuition runs deep and that it is driven by people's tendency to view a reductive, mechanistic explanation of behavior—for instance, in the neuroscientific language of neural processes and chemical reactions—as inconsistent with a mentalistic (or intentional) explanation—in the psychological language of thoughts, desires, and plans. Because people also tend to ascribe free will (FW) and moral responsibility (MR) only to agents whose actions can be understood in terms of their mental states, people tend to see reductive mechanism as incompatible with FW and MR. That is, we think that most ordinary folk have intuitions about freedom and responsibility that accord with a position we will call *Mechanism Incompatibilism* (MI): FW and MR are incompatible with reductive mechanism.

We take it to be a separate question whether most people express intuitions that accord with the position at the heart of philosophical debates about free will, what we will call *Pure Incompatibilism* (PI), the view that *determinism* is incompatible with FW and MR. Determinism is a thesis about the relations (perhaps causal relations) between states of a system at different times as governed by the laws of nature. It says that, holding fixed the laws of nature and the state of the system at one time, there is only one possible state of the system at any other time.[1] Given this definition, it is possible for a system to be deterministic even though reductive, mechanistic explanations are *not* sufficient to describe it. For instance, a system with irreducible psychological laws—that cannot be explained solely in terms of the causal interactions among lower-level mechanisms—may still be deterministic. Conversely, reductive mechanism (hereafter "mechanism") is entirely consistent with the falsity of determinism. For instance, all the causal interactions in our universe may ultimately be reducible to quantum events, even if quantum mechanics is indeterministic. We mention these examples to emphasize that determinism and mechanism are distinct metaphysical theses: Whether, for instance, human beings can be understood as mechanistic systems is a different question from whether our universe is deterministic (or whether human behavior obeys deterministic laws).

And this distinction entails that MI is a distinct position from PI. One could, for instance, hold that MI is true but PI is false. Indeed, we think it is likely that *most* people have intuitions that indicate an acceptance of MI but *not* of PI. If this is true, it would be easily overlooked, since most people—including many scientists and journalists who write about free will—conflate determinism with mechanism (which helps to explain why they suggest, as in the quotations above, that by showing that humans are mechanistic systems, neuroscience also proves that free will is an illusion). As we will explain below, this conflation of mechanism with determinism would also provide an explanation for the common claim that PI represents the commonsensical and intuitive position, a claim we think is false.

One way to examine the relevance of determinism and mechanism to people's judgments about FW and MR is to present them with agents in a range of scenarios, varied according to whether determinism and/or mechanism holds in the

1. See, for example, van Inwagen (1983)—for his definition, the system discussed is the entire universe. Some definitions of determinism stipulate that the relation holds only from past states to future states.

scenario, and examine their judgments about the agents' freedom and responsibility. The experiments we describe in this paper were developed to test folk intuitions about FW and MR in this way. But before we describe the experiments and results, we will briefly explain why we think such information about folk intuitions about FW and MR is relevant to the philosophical debates about these issues.

1. WHY MAP ORDINARY INTUITIONS?

The traditional philosophical debates about free will take it that the most natural and important potential challenge to people's belief in freedom and responsibility is determinism. As Galen Strawson (1986) puts it, it is "in our nature to take determinism to pose a serious problem for our notions of responsibility and freedom" (89). The debate then focuses primarily on the question of whether it is true that determinism is in fact incompatible with FW and MR. Those who believe it is, the *incompatibilists*, often motivate their position by appealing to the "natural intuitiveness" of their position, claiming that *ceteris paribus*, we should favor incompatibilism because "most ordinary persons start out as natural incompatibilists" (Kane 1999, 217), and "we come to the table, nearly all of us, as pretheoretic incompatibilists" (Ekstrom 2002, 310).[2] We will argue that the focus of this debate is misplaced. Most people do not initially find determinism *per se* to undermine FW and MR. Rather, PI merely *appears* to be the intuitive position because it trades on the intuitions that support MI by conflating determinism with mechanism. At a minimum, the MI intuition is more central to ordinary people's understanding of the relevant concepts and practices, and, we believe, more significant given current scientific discussions of human nature.

But why should philosophers engaged in the debate care about what ordinary people think about such complicated philosophical concepts and debates? What bearing can information about folk intuitions have on these debates? These questions have been addressed in more detail elsewhere (see, e.g., Nadelhoffer and Nahmias 2007), and, as the essays in this volume attest, the answers to these questions will also depend on one's meta-philosophical presuppositions and on the nature of the philosophical debate at issue. For now, we will simply point out that in debates about FW and MR, philosophers on all sides of the debate regularly refer to ordinary intuitions and common sense. In addition to the incompatibilists mentioned above, compatibilist Susan Wolf (1990) illustrates this practice when she says of her view that it "seems to accord with and account for the whole set of our intuitions about responsibility better than . . . the leading alternatives" (89). Presumably, philosophers make such appeals to ordinary intuitions because this debate (perhaps unlike some more abstruse philosophical debates) is about issues that are of great significance to ordinary people, intimately connected to our beliefs about morality and the meaning of life, and to the practices of praising and blaming, punishing and rewarding, not to mention associated legal practices. Philosophical debates about free will, metaphysical as they may be, should thus be relevant to people outside philosophy. Conversely, what people think about FW

2. For more examples see Nahmias et al. (2006).

and MR should also matter to the philosophical debates. If these debates end up disputing philosophical analyses of concepts that no longer connect with the folk concepts and practices, then these debates risk being irrelevant. And at best, such philosophical theories might need to *revise* the folk's mistaken views (see Vargas 2005), but this move still requires understanding both what the folk's beliefs actually are and what cognitive processes produce their beliefs so that such a revision might be feasible.

In any case, mapping the relevant beliefs and intuitions in a systematic, empirically informed way can play a useful role in the development of a viable theory of FW and MR, especially since philosophers tend to offer, from the armchair, conflicting claims about what is intuitive and what our ordinary practices and beliefs are. If such a mapping of intuitions shows that people have vague or conflicting intuitions, folk concepts and theories will require clarification, perhaps even revision. But even in the case of revision, we should follow David Lewis's (1986) advice regarding philosophical theorizing: "[We are] trying to accomplish two things that somewhat conflict. We are trying to *improve* that theory, that is to change it. But we are trying to improve *that* theory, that is to leave it recognizably the same theory we had before" (134). Any revision of our ordinary understanding of FW and MR has to be something that ordinary people can understand as a theory of *those* concepts and can employ in the context of *those* practices. Furthermore, the revisionist should offer a plausible explanation for why people hold any purportedly mistaken intuitions and perhaps also for how people's mistaken beliefs can be effectively revised.

Indeed, by mapping folk intuitions about FW and MR, we can better understand the cognitive processes that generate these intuitions. And this will put philosophers in a better position to provide plausible explanations of why certain views are attractive even if they are mistaken—and to understand how mistaken intuitions might be corrected. The best way to begin such projects is to utilize the methods of empirical psychology; for instance, present people with different scenarios, varied in systematic ways, in order to examine which variables correlate with differing responses. Under the rubric of "experimental philosophy," this method has been used by philosophers and psychologists to study people's intuitions about various philosophical issues (see Nadelhoffer and Nahmias 2007, for references). We have adopted this method for our current project.

We want to emphasize that we do not think that experimental evidence about folk intuitions can provide definitive answers to these philosophical debates. Rather, the evidence is best seen as ammunition that may be used to advance a philosophical position. Here, our primary goal was to develop experiments to test people's intuitions about FW and MR in order to better understand the underlying psychological processes that generate such intuitions. We will then use the data from these experiments to advance the philosophical conclusion that debates about FW and MR should shift their attention away from questions about determinism toward questions about the relationship between mentalistic explanations and mechanistic explanations. To motivate this project, we will briefly discuss some relevant work in experimental philosophy on these issues.

2. A BRIEF HISTORY

As we have seen, competing philosophers have made competing claims about people's intuitions about freedom and responsibility. But they make these claims based on their own intuitions or speculation about the intuitions of other people (e.g., their students). Nahmias, Morris, Nadelhoffer, and Turner (2005, 2006) tested these claims about folk judgments of FW and MR in a more systematic way. Using three different descriptions of determinism, they found that a significant majority of participants judged that agents in a deterministic scenario act of their own free will and are morally responsible. These results seem to indicate that, contra the common claim that compatibilism is counterintuitive, most people are in fact "natural *compatibilists*." These results call for an explanation for why so many philosophers have assumed that determinism is intuitively threatening to FW and MR. Below, we will argue for, and present evidence in support of, an explanation of the *apparent* intuitive appeal of incompatibilism. But we should first canvass some responses incompatibilists have available to offer to such evidence.

First, they can argue that the experimental results have not accurately uncovered people's *actual* intuitions about these issues. We will leave such methodological criticisms to be developed—and preferably empirically tested—while noting that they will have to find flaws with the study presented here as well. Second, they might argue that such experiments have not uncovered people's *relevant* intuitions about these issues; the relevant intuitions are about deliberation, choice, ability, etc., and it is these intuitions that commit people to incompatibilist conclusions. For instance, one might argue that people intuitively conceive of FW and MR as requiring *alternatives* that are incompatible with the truth of determinism or that people will take the premises and principles of an incompatibilist argument to be highly intuitive. Of course, many compatibilists argue that these conceptions are not intuitive or that the premises and principles in such arguments are either false, invalid, or question begging (for instance, some argue that the type of alternatives determinism precludes are *not* intuitively required for FW or MR). Of course, at this point, philosophical argument ensues and perhaps folk intuitions should get left behind. But if so, incompatibilists' appeals to commonsense intuitions to advance their position or to shift the burden of proof to compatibilists would be infelicitous (see Nahmias et al. 2006).

From the perspective of experimental philosophy, the relevant work here might include testing people's intuitions about the *premises* of these arguments,[3] or testing their intuitions before and after being exposed to the *arguments* themselves, perhaps comparing a group who are presented the arguments by an incompatibilist teacher with a group instructed by a compatibilist teacher. For a *critic* of experimental philosophy, the conclusion might be that folk intuitions are unreliable or simply irrelevant to the philosophical debate—only philosophically informed intuitions should be considered. But we challenge such a critic to explain (1) why, if folk intuitions are considered unreliable, competing philosophers' conflicting intuitions *should* be considered reliable; and (2) how, if we abandon ordinary intuitions and

3. Along with Tamler Sommers, Nahmias is working on such a project.

conceptual usage, we can maintain the connection between philosophical discussions of concepts such as FW and MR and the ordinary concepts and practices that people actually care about (as discussed above) or figure out how to revise these concepts and practices.

Without giving up reference to ordinary intuitions, another way to respond to results indicating that the folk have compatibilist intuitions is to explain why people express intuitions that might not reflect their considered beliefs or theories. In part to offer such an explanation for people's compatibilist judgments, Shaun Nichols and Joshua Knobe (forthcoming) developed experiments aimed to explore the psychological mechanisms that generate intuitions about moral responsibility. In their studies, "one condition, the *concrete* condition, was designed to elicit greater affective response; the other condition, the *abstract* condition, was designed to trigger abstract, theoretical cognition" (9). They found that in the concrete condition, when people were presented with a deterministic universe in which a specific agent committed a morally reprehensible act (killing his wife and children), most judged him to be "fully morally responsible." But they also found that in the abstract condition, when asked whether it is possible for a person to be fully morally responsible in the deterministic universe, the vast majority of people responded "no." Nichols and Knobe suggest that the disparity between folk responses in these concrete and abstract cases has to do with the psychological mechanisms driving people's intuitions. They claim that the immoral action in the concrete case engages people affectively in a way that leads them to offer compatibilist judgments. Moreover, they suggest a "performance error model" for these judgments: "[T]he normal competence with responsibility attribution is skewed by the emotions; that explains why there is such a large difference between the high and low affect in the determinist conditions" (23–24). They suggest that this model provides an explanation for the compatibilist results found by Nahmias et al. (2005, 2006).

We are not yet convinced that Nichols and Knobe's results suggest that, in the scenarios they use, people are expressing *incompatibilist* intuitions. We worry that their experimental setup does not test whether people find determinism *per se* incompatible with MR—rather, we think they have found evidence that people judge in the abstract case that "full moral responsibility" is impossible in a world in which decisions *have to happen* the way they do (see Nahmias 2006; Turner and Nahmias 2006).[4] Nonetheless, we are convinced that Nichols and Knobe have found impressive evidence that people's judgments about MR are significantly influenced by their emotional responses to agents performing bad acts. In our current study, we attempted to replicate Nichols and Knobe's finding, but we will suggest another way to interpret the effect, and we will question whether the performance error model is the best way to interpret the influence of affect on people's judgments of responsibility. We will also follow their strategy of develop-

4. Nichols and Knobe's scenarios end by contrasting the "deterministic" universe (A) with the "indeterministic" universe (B): "The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision—given the past, each decision *has to happen* the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision *does not have to happen* the way that it does."

ing experiments to examine the psychological processes that generate folk intuitions about MR—but we will propose a model that offers a way to explain why people may appear to express *incompatibilist* intuitions.

3. WHICH INCOMPATIBILISM IS INTUITIVE?

Incompatibilists often claim that having free will is something like being in a "garden of forking paths," but that the truth of determinism would entail that the garden is an illusion—that there is only one path even though it may *look* to its travelers as if it forks in multiple directions. Carl Ginet explores this possibility using a disconcerting illustration of what he takes determinism to suggest:

> [I]n those rides that amusement parks sometimes provide, in which one sits in a car that follows a track through some darkened room of illuminated objects, the car sometimes has a steering wheel. If one turns the wheel in the directions suggested by the environment—directions in which the car is actually going—one can easily get the feel that one is steering the car—even though one knows all along that he is not. A child might think that he actually was steering the car. (Ginet in Fischer 1994, 14)

Ginet's picture suggests that determinism entails that our sense of control is, like the child's, illusory because we are pushed along "predestined grooves" (Feinberg in Fischer 1994, 5). If determinism is true, then our conscious mental processes that we think steer us in our decision-making are supposedly epiphenomenal. As Feinberg puts it, "our self-monitoring and self-critical capacities, so essential to human nature, might as well dry up and wither; they would no longer have any function" (5). The threat to FW and MR here seems focused *not* on the fact that there are sufficient conditions, going back to before we were born, to ensure our decisions and actions, but rather on the purported implication that our deliberations and conscious purposes are *bypassed* by forces that are out of our control.

We think that determinism *per se* does not suggest this bypassing threat, since determinism does not imply that our deliberations and conscious purposes are causally irrelevant to what we do (at least, demonstrating such an implication would require a convincing argument). But *without* presenting determinism in such a way that it suggests such bypassing, we do not think most people would find it intuitively threatening to FW and MR. Rather, people are likely to see such a threat only if their decision-making is described in reductive, mechanistic terms. Again, the idea is that most people have intuitions that support Mechanism Incompatibilism rather than Pure Incompatibilism.[5]

There is evidence that people employ different cognitive processes when they are explaining systems in terms of mechanistic causal relations and when they are explaining systems by ascribing mental states to the system—the latter is sometimes discussed in terms of theory of mind or "mindreading" abilities (e.g.,

5. For initial evidence in support of this claim, see Monterosso, Royzman, and Schwartz (2005) and Nahmias (2006).

Bloom 2004). To adopt the language of Daniel Dennett (1973), when people take a "mechanistic stance" toward a system, this tends to displace "any explanation in terms of beliefs, desires, and intentions" (151)—that is, explanation in terms of what he calls the "intentional stance" (Dennett 1987).

Now, we agree with Dennett that mechanistic systems (such as humans) can also be purposive and intentional systems, contra philosophers such as Norman Malcolm (1968), who suggests that "purposive explanations . . . would be refuted by the verification of a comprehensive neurophysiological theory of behavior" (132). However, we think it is nonetheless highly intuitive to view reductive mechanistic explanations as conflicting with and superseding mentalistic explanations (involving beliefs, desires, intentions, etc.). And it is *this* intuition—that mechanism conflicts with mentalism—that underlies people's MI intuitions, since people (rightly) assume that free and responsible behavior must be caused by the right sorts of mental processes and states (e.g., the agent's deliberations, desires, goals, and reasons). Indeed, it is typically the compatibilist who emphasizes this condition for FW and MR. And we think this MI intuition is then easily mistaken as support for PI such that philosophers and others take it as obvious and intuitive to ordinary folk that freedom and responsibility are only possible if determinism is false.[6]

Our theory generates several predictions:

(1) If presented with an agent in a deterministic world whose decision-making is described in mechanistic terms (e.g., neural processes and chemical reactions in the agent's brain), most people will judge that the agent does *not* have FW and is *not* MR.
(2) If presented with an agent in a deterministic world whose decision-making is described in *non*mechanistic, psychological terms (e.g., thoughts, desires, and plans in the agent's mind), most people will judge that the agent *does* have FW and *is* MR.
(3) If presented with an agent in an *in*deterministic world whose decision-making is described in mechanistic terms, most people will judge that the agent does *not* have FW and is *not* MR.
(4) If presented with an agent in an *in*deterministic world whose decision-making is described in *non*mechanistic, psychological terms, most people will judge that the agent *does* have FW and *is* MR.

Evidence in support of these intuitions would suggest that most people do not have PI intuitions; instead, they have MI intuitions. For them, the truth of determinism or indeterminism is not the issue most relevant to questions about free will and moral responsibility. Rather, whether mechanism is true is the issue most people find relevant to these questions. It is with the first two of these predictions in mind that we generated the scenarios for our study.

6. Notice that this discussion might also be put in terms of *reductionistic* explanations of human behavior. Kim's (1998) exclusion argument against mental causation might be seen as formalizing the intuition that a complete causal explanation of human behavior at the neurobiological level supersedes causal explanation at the psychological level.

4. THE SCENARIOS

While the earlier research of Nahmias et al. (2005, 2006) focused on the *content* of folk intuitions about judgments of FW and MR, our current interest lies primarily in discovering features of the cognitive processes that drive these judgments. As discussed above, Nichols and Knobe have provided some insight into these processes—high affect increases judgments of MR. We believe that new scenarios can provide additional insight into the relevant cognitive processes. Whereas high affect may drive certain intuitions about FW and MR, we believe that the way determinism is described can also influence these intuitions. If a scenario presents determinism in a way that leads people to adopt the mechanistic stance toward an agent, then people will be less inclined to attribute FW and MR toward the agent. Conversely, if determinism is presented such that people interpret an agent's conscious deliberations and reflectively held values and goals (or in more folksy terms, thoughts, desires, and plans) to be causally efficacious, then the relevant cognitive processes will tend to generate attributions of FW and MR.

When people view another agent from the mechanistic stance, they are likely to take what Peter Strawson (1962) calls *objective attitudes* toward the agent. The agent is treated as an object to "be managed or handled or cured or trained" (66); we do not engage the participant reactive attitudes—"natural human reactions to the good or ill will or indifference of others towards us, as displayed in *their* attitudes and actions" (67). If people don't see the agent's conscious deliberations and reflectively held values and goals as playing a causal role in subsequent behavior, if the behavior is "unintelligible to us, perhaps even to him, in terms of conscious purposes" (73), then we do not see the agent as an appropriate member of the moral community; people do not take the "participant stance" toward the agent. We think that presentations of determinism that evoke reductive mechanism will engage this feature of our cognitive processing.

We thus developed scenarios that describe determinism but vary according to whether agents' decision-making is described in terms of neuroscientific, mechanistic processes ("Neuro scenarios") or in terms of psychological, intentional processes ("Psych scenarios"). It is a challenge to describe the philosophical thesis of determinism in a way that makes it both understandable to non-philosophers and precise enough to properly test their intuitions about its compatibility with FW and MR while also being careful to avoid begging any questions about whether determinism entails, for instance, that decisions *have to happen* the way they do. We described deterministic worlds in which agents' decisions are (1) completely caused by prior events, and (2) those prior events were completely caused by earlier events going back to events before the agent was born, such that (3) the prior events will definitely cause the later events.[7] In the first set of scenarios we developed, we described an alternate universe, Erta, "that is very similar to our

7. We recognize that some people may be unsatisfied by the way we chose to describe determinism, and we hope such critics will consider developing other formulations of determinism for future research.

world," in part because we thought it might help people reason counterfactually about the implications of determinism, especially if they do *not* think our universe is deterministic. Participants were asked to answer questions (e.g., about FW and MR) based on what the story said about the alternate world: "Even if you don't think that what it says is true in our world, imagine that it is true in Erta." The first two scenarios look like this (with the variations for the Psych scenarios in brackets):

> Story: On Erta, the landscape and life are very similar to Earth, and there are advanced life forms called Ertans who look, talk, and behave very much like we do. For instance, the Ertans have families, schools, various jobs, parties, arguments, etc. However, the Ertans' science has advanced far beyond ours. Specifically, Ertan neuroscientists [psychologists] have discovered exactly how Ertans' brains [minds] work. For instance, they have discovered that whenever an Ertan is trying to decide what to do, the decision the Ertan ends up making is completely caused by the specific chemical reactions and neural processes [thoughts, desires, and plans] occurring in his or her brain [mind]. The neuroscientists [psychologists] have also discovered that these chemical reactions and neural processes [thoughts, desires, and plans] are completely caused by the Ertan's current situation and the earlier events in his or her life. These earlier events were also completely caused by even earlier events, eventually going all the way back to events that occurred before the Ertan was born.
>
> So, once specific earlier events have occurred in an Ertan's life, these events will definitely cause specific later events to occur. For instance, once specific chemical reactions and neural processes [thoughts, desires, and plans] occur in the Ertan's brain [mind], they will definitely cause the Ertan to make the specific decision he or she makes.

These scenarios describe the Ertan agents in an abstract way, but we were also interested in replicating and examining Nichols and Knobe's finding that people's judgments are significantly influenced by the presentation of concrete agents performing morally salient actions. To that end, we developed four more scenarios, two that describe a specific agent doing a morally bad action ("Neuro Bad" and "Psych Bad") and two that describe an agent doing a morally good action ("Neuro Good" and "Psych Good"). For these scenarios, the final paragraph (above) was replaced with this paragraph (variations in brackets):

> So, once specific earlier events have occurred in an Ertan's life, these events will definitely cause specific later events to occur. For example, one day an Ertan named Smit decides to kill his wife so that he can marry his lover [donate a large sum of money to an orphanage in his community], and he does it. Once the specific chemical reactions and neural processes [thoughts, desires, and plans] occur in Smit's brain [mind], they will definitely cause his decision to kill his wife [donate a large sum of money to an orphanage in his community].

Finally, we developed two scenarios that described the *real* world as deterministic, one using neuroscientific language and one using psychological language ("Neuro Real" and "Psych Real"), in part to examine interesting differences found by Nichols and Roskies (unpublished) in people's responses to agents in real versus alternate worlds. For these scenarios, participants were asked to judge what they think about people's FW, MR, etc. *if* the neuroscientists/psychologists are right about what they believe according to this description (Psych variations in brackets):

> Most respected neuroscientists [psychologists] are convinced that eventually we will figure out exactly how all of our decisions and actions are entirely caused. For instance, they think that whenever we are trying to decide what to do, the decision we end up making is completely caused by the specific chemical reactions and neural processes [thoughts, desires, and plans] occurring in our brains [minds]. The neuroscientists [psychologists] are also convinced that these chemical reactions and neural processes [thoughts, desires, and plans] are completely caused by our current situation and the earlier events in our lives, and that these earlier events were also completely caused by even earlier events, eventually going all the way back to events that occurred before we were born.
>
> So, if these neuroscientists [psychologists] are right, then once specific earlier events have occurred in a person's life, these events will definitely cause specific later events to occur. For instance, once specific chemical reactions and neural processes [thoughts, desires, and plans] occur in the person's brain [mind], they will definitely cause the person to make the specific decision he or she makes.

We ran these real world variations only in the abstract condition (not "Real Bad" and "Real Good" variations), so we ran a total of eight different scenarios to test our main hypotheses.[8] In the next two sections, we will discuss in some detail the methods and results of our study. Some readers may not be interested in these details, but given that our conclusions rest heavily on experimental data, we think it is important to explain how we collected and analyzed that data.[9]

8. Additionally, we examined some other issues relevant to FW and MR, which we discuss in the Appendix.

9. Experimental philosophers have been challenged (e.g., by Bernstein, unpublished) to ensure that their methodology accords with the standards of empirical psychology. We have attempted to live up to these standards, and we seek feedback on whether we have succeeded. We cannot include here *all* of the details about our rather ambitious experiment, which used numerous scenarios (8), experimental questions (20), and participants (over 1,100). We also hope to run some of the other possible variations, such as the Real Bad and Real Good scenarios and some variations using descriptions of *in*determinism.

5. METHODS

For this study, 1,124 Georgia State University undergraduates participated for extra credit in a critical thinking course. Participants were contacted through e-mail correspondence and randomly assigned to one of the eight versions of the experimental task. Participants were 41.2 percent male and 57.5 percent female, with 1.3 percent declining to give gender information.

We used online software from QuestionPro to develop surveys online.[10] Using a $2 \times 3 \times 2$ between-participants design, eight scenarios were generated by systematically varying (1) whether the universe described was our world or a similar alternate world [Real versus Erta]; (2) whether the agents were described abstractly or whether a particular agent was described performing a concrete bad act (killing his wife) or good act (giving to charity) [Abstract versus Bad versus Good]; and (3) whether the determinism in the scenario was described in neuroscientific language or psychological language [Neuro versus Psych].

Participants began by reading a general description of the task, providing informed consent, and then reading one of the eight scenarios. After reading the scenario, they answered a manipulation check to test that they understood the deterministic nature of the scenarios. Then they answered ten experimental questions designed to probe their intuitions about FW and MR (answers to the first five questions provided the variables of most interest to us). These questions were randomized to decrease the likelihood of order effects. Participants responded to each question by indicating their level of agreement using a 6-point rating scale (strongly disagree, disagree, somewhat disagree, somewhat agree, agree, strongly agree, plus "I don't know"). Specifically, participants were asked to rate, based on the scenario they just read, the degree to which they agreed or disagreed with the following statements (below are the questions from the Ertan Abstract scenarios, with representative samples of variations from some of the other scenarios in brackets):

(1) Ertans' decisions are up to them. [Bad: Smit's decision to kill his wife was up to him.]
(2) Ertans are able to make decisions of their own free will. [Neuro Real: If the neuroscientists are right, then people make decisions of their own free will.]
(3) Ertans should be held morally responsible for their decisions. [Psych Real: If the psychologists are right, then people should be held morally responsible for their decisions.]
(4) If an Ertan does something bad, then he or she deserves to be blamed for it. [Bad: Smit deserves to be blamed for killing his wife.]

10. Although some have voiced worries about the reliability of data collected from online research, several studies have contrasted data from online research with data collected through more traditional methods and consistently found results to be identical (see Birnbaum 2000). We viewed the many advantages of taking a Web-based approach—economy and ease of data collection as well as potential for increased internal validity—to far outweigh the disadvantages.

(5) If an Ertan does something good, then he or she deserves to be praised for it. [Good: Smit deserves to be praised for donating money to the orphanage.]
(6) Smit controls his decision to donate money to the orphanage.
(7) Ertans deliberate about what to do.
(8) Ertans are able to have deep personal relationships (e.g., friendship and love).
(9) Ertans are able to lead meaningful lives. [Neuro Real: If almost everyone came to believe what the neuroscientists believe, people would still be able to lead meaningful lives.]
(10) Ertans are likely to act as ethically as we do.

After providing responses to these questions, participants were asked to briefly explain their responses to questions 2, 3, and 9 above (questions about FW, MR, and meaningful lives). Participants then answered a second manipulation check to ensure that they had kept the deterministic nature of the scenario in mind while responding. They then answered several further questions about: (11) whether the description in the scenario accurately describes human decision-making, (12/13) whether, given an agent's neural/psychological states, the agent might (or "can" in some scenarios) make more than one decision, (14) whether retributive punishment is appropriate and (15) whether retributive punishment would be appropriate *given* the facts in the scenario, (16) whether humans have nonphysical souls, (17) whether humans make decisions of their own free will, and (18) whether humans have free will *only* because they have nonphysical souls. (See Appendix for discussion).

Finally, participants completed the survey by answering several demographic questions (gender, age, religious affiliation and religiosity, and whether they had taken a college philosophy course) and by reading a short debriefing form.

6. MAIN RESULTS

We excluded from analysis (and from results presented in Table 1) all participants who did not complete the survey in full and all participants who responded incorrectly to either of the two manipulation checks (twenty-two percent of all participants).[11] Since we are interested in *pre*-philosophical folk intuitions, we also excluded those participants who indicated that they had taken a college philosophy class (twenty-nine percent overall). It is an interesting question whether prior philosophical training influences people's intuitions about free will, determinism, and moral responsibility. We suspect that such training would sometimes influence such intuitions, but probably only after significant exposure to the relevant arguments (and it is unclear whether such exposure would change, strengthen, or weaken the intuitions). Since most of our participants who indicated they had taken a philosophy course had likely taken only one introduction to philosophy

11. Although this is a large group to exclude from our final sample, given the high demand characteristics of the task, we expected a significant loss of participants due to these manipulation checks.

Table 1. Descriptive Results

Questions	Real world (A–B)		Ertan (alternate) world (C–H)					
	Abstract		Abstract		Concrete bad		Concrete good	
	A. Neuro Real $n = 81$	B. Psych Real $n = 71$	C. Neuro Ertan $n = 90$	D. Psych Ertan $n = 65$	E. Neuro Bad $n = 108$	F. Psych Bad $n = 111$	G. Neuro Good $n = 48$	H. Psych Good $n = 58$
1. Decisions "up to" agent	34.3	85.9	40.4	56.3	61.7	72.7	54.2	64.3
2. Free will	38.3	82.9	39.1	53.1	59.8	66.1	57.4	61.1
3. Morally responsible	40.7	88.6	52.4	71.9	79.2	81.1	63.0	68.5
4. Deserve blame	37.7	85.7	50.6	70.3	74.3	85.6	NA	NA
5. Deserve praise	48.7	85.9	67.0	78.1	NA	NA	70.5	75
6. Control	38.8	80.3	NA	NA	NA	NA	52.1	54.4
7. Deliberate	57.5	85.5	59.5	64.5	52.8	49.5	54.5	56.4
8. Relationships	60.8	87.7	81.4	83.6	81.1	79.2	86.7	81.5
9. Meaningful lives	62.0	85.1	74.4	88.7	75.5	75.5	80.9	81
10. Act as ethically	53.2	75.4	62.7	71.0	57.3	52.8	68.9	69.6
11. Description is true of us	38.3	74.6	45.6	53.8	47.2	50.5	33.3	60.3
12. *Might* make only one decision	NA	NA	67.8	73.8	71.3	70.3	NA	NA
13. *Can* make only one decision	72.8	46.5	NA	NA	NA	NA	70.8	58.6
14. Criminals deserve to suffer	26.9	42.3	51.1	46.2	49.1	53.2	41.7	55.2
	34.6	25.4	26.7	35.4	28.7	27.0	27.1	29.3
	38.5	32.4	21.1	18.5	22.2	19.8	29.2	15.5
15. Criminals in *scenario* deserve to suffer	NA	NA	14.4	32.3	33.3	37.8	35.4	34.5
			55.6	47.7	38	31.5	29.2	37.9
			30.0	20.0	28.7	30.6	35.4	27.6
16. Humans have nonphysical souls	63.0	63.4	74.4	70.8	66.7	66.7	47.9	55.2
	16.0	22.5	14.4	15.4	20.4	16.2	20.8	24.1
	21.0	14.1	11.1	13.8	13	17.1	31.3	20.7
17. Humans have free will	80.2	85.9	93.3	89.2	94.4	91	93.8	93.1
	16	9.9	5.6	10.8	5.6	8.1	4.2	6.9
	3.7	4.2	1.1	0	0	0.9	2.1	0
18. Free will *only* because of souls	24.7	16.9	NA	NA	NA	NA	14.6	22.4
	33.3	46.5					47.9	44.8
	42.0	36.6					37.5	32.8

Note: For questions 1–13, these descriptive results are presented as the *percentage of participants who agreed* (somewhat too strongly). For these questions, the few participants (almost always fewer than five percent) who answered "I don't know" are not included, so the remaining percentage indicates those participants who disagreed (somewhat too strongly). For questions 14–18, results are presented as the percentage of participants who agreed versus the percentage who disagreed versus the percentage who answered "I don't know" (a much higher proportion on most of these questions). NA = Not asked or not analyzed.

course, we predicted that their intuitions would *not* vary significantly from the general sample. Indeed, for all the comparisons we ran, we found no significant differences between groups (in only one sample set, the Psych Real scenario, did we find any statistical trends based on prior philosophy courses). Nonetheless, for the results presented here, we excluded participants who had taken a prior philosophy course. All analyses were conducted on the remaining sample of 632 participants.

A 2 (mechanism: Psych, Neuro) × 2 (universe: Real, Erta) Analysis of Variance (ANOVA) was run on the mean responses to our five primary variables of interest ("up to," FW, MR, blame, and praise; questions 1–5).[12] For judgments of each of the five variables, we found similar results that indicated highly significant differences ($p < .001$) depending on whether agents' decision-making in the scenario was described in psychological terms or neuroscientific (mechanistic) terms. Each ANOVA showed a significant main effect for mechanism (Psych versus Neuro),[13] no main effect for universe,[14] and a significant interaction effect[15] (See Figure 1).

Because the interaction effect made it difficult to interpret the significant main effect of mechanism, we ran additional *t*-tests comparing the mean responses to each of the five variables in the Neuro Real and Psych Real scenarios, as well as in the Neuro Ertan and Psych Ertan scenarios. For each of the five questions, we found that mean responses in the Neuro Real scenarios were consistently lower on the scale of disagreement/agreement than mean responses in the Psych Real scenarios. In each case, these differences were highly significant ($p < .001$).[16] When comparing the Ertan scenarios, we again found that responses in the Neuro scenarios were lower than responses in the Psych scenario for each variable. With the exception of judgments concerning praise, the differences between the Neuro Ertan and Psych Ertan scenarios were significant, although not nearly to the degree of those in the Real world comparisons.[17] We also ran *t*-tests comparing mean responses to the Neuro Real and Neuro Ertan, as well as Psych Real and Psych Ertan conditions. For each of the five variables, mean responses in the Neuro Ertan condition were *higher* than in the Neuro Real scenarios but only reached signifi-

12. Means and standard deviations (SDs) for *Neuro Real*: (Up To: M = 2.94, SD = 1.54; FW: M = 2.89, SD = 1.54; MR: M = 3.15, SD = 1.45; Blame: M = 3.01, SD = 1.44; and Praise: M = 3.36, SD = 1.55), *Psych Real* (Up To: M = 4.46, SD = 1.27; FW: M = 4.39, SD = 1.37; MR: M = 4.79, SD = 1.31; Blame: M = 4.43, SD = 1.27; and Praise: M = 4.44, SD = 1.1), *Neuro Ertan*: (Up To: M = 3.18, SD = 1.38; FW: M = 3.16, SD = 1.45; MR: M = 3.64, SD = 1.36; Blame: M = 3.59, SD = 1.29, and Praise: M = 4.05, SD = 1.33), and *Psych Ertan*: (Up To: M = 3.63, SD = 1.44; FW: M = 3.70, SD = 1.45; MR: M = 4.31, SD = 1.32; Blame: M = 4.22, SD = 1.20; and Praise: M = 4.30, SD = 1.22).

13. Up To: $F(1, 301) = 36.45$, $p < .001$; FW: $F(1, 298) = 36.375$, $p < .001$; MR: $F(1, 295) = 52.81$, $p < .001$; Blame: $F(1, 294) = 44.973$, $p < .001$; and Praise: $F(1, 297) = 18.84$, $p < .001$.

14. Up To: $F(1, 301) = 3.356$, $p = .068$; FW: $F(1, 298) = 1.475$, $p = .226$; MR: $F(1, 295) = .005$, $p = .946$; Blame: $F(1, 294) = 1.416$, $p = .235$; and Praise: $F(1, 297) = 3.189$, $p = .075$.

15. Up To: $F(1, 301) = 10.96$, $p = .001$; FW: $F(1, 298) = 7.973$, $p = .005$; MR: $F(1, 295) = 9.294$, $p = .003$; Blame: $F(1, 294) = 6.574$, $p = .011$; Praise: $F(1, 297) = 7.28$, $p = .007$.

16. Up To: $t(150) = 6.59$; FW: $t(149) = 7.243$; MR: $t(149) = 7.243$; Blame: $t(145) = 6.28$; and Praise: $t(147) = 4.853$. All results significant at $p < .001$.

17. Up To: $t(151) = 1.93$, $p = .05$; FW: $t(149) = 2.26$, $p = .025$; MR: $t(146) = 3.01$, $p = .003$; and Blame: $t(149) = 3.065$, $p = .003$.

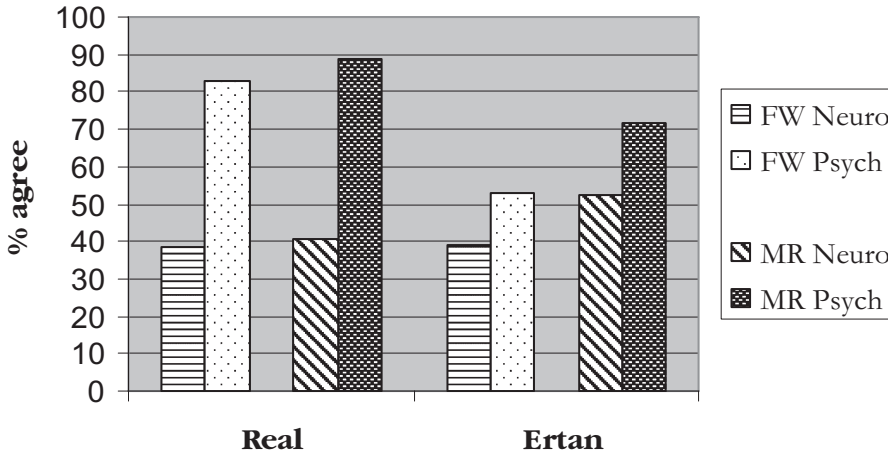

Figure 1. The effect of mechanism.
Percentage of participants who agreed that agents make decisions of their own free will (FW) and are morally responsible (MR) for their decisions.

cance in the MR, praise, and blame conditions.[18] Conversely, in the Psych Ertan and Psych Real comparisons, mean responses to all variables were *lower* in the Ertan scenario than in the Real world scenario, although only "up to," FW, and MR questions reached significance.[19]

Based on these analyses, it appears that regardless of whether a deterministic scenario is set in our own world or in an alternate world, how it is described—either in neuroscientific or psychological language—has a significant impact on judgments of FW, MR, and related concepts. But interestingly, when the situation is set in our world as opposed to an alternate world, people's responses indicate more pronounced intuitions, particularly in the Psych scenarios. For each of the five variables, mean responses in the Neuro *Real* scenario were *lower* than in the Neuro *Ertan* scenario, whereas mean responses in the Psych *Real* scenario were much *higher* than in the Psych *Ertan* scenario. That is, people offered particularly compatibilist-friendly responses in the Psych Real scenarios. And the threat of mechanism to FW and MR was manifested most clearly in the Neuro Real scenarios. A chi-square test indicates that these responses were significantly different from chance for all but two of the variables in these scenarios.[20]

18. Up To: $t(168) = 1.08$, $p = .283$; FW: $t(166) = 1.18$, $p = .241$; MR: $t(163) = 2.26$, $p = .025$; Blame: $t(162) = 2.68$, $p = .008$; Praise: $t(164) = 3.048$, $p = .003$. See note 12 for means and SDs.

19. Up To: $t(133) = 3.592$, $p < .001$; FW: $t(132) = 2.802$, $p = .006$; MR: $t(132) = 2.084$, $p = .039$; Blame: $t(133) = .699$, $p = .486$; Praise: $t(132) = .981$, $p = .486$. See note 12 for means and SDs.

20. Neuro Real Up To: $\chi^2(1,n = 81) = 7.716$, $p = .005$; FW: $\chi^2(1,n = 81) = 4.457$, $p = .035$; MR: $\chi^2(1,n = 81) = 2.778$, $p = .096$; Blame: $\chi^2(1,n = 77) = 4.688$, $p = .03$; Praise: $\chi^2(1,n = 78) = .051$, $p = .821$ Psych Real Up To: $\chi^2(1,n = 71) = 36.634$, $p < .001$; FW: $\chi^2(1,n = 70) = 30.229$, $p < .001$; MR: $\chi^2(1,n = 70) = 41.657$, $p < .001$; Blame: $\chi^2(1,n = 70) = 35.714$, $p < .001$; Praise: $\chi^2(1,n = 71) = 36.634$, $p < .001$.

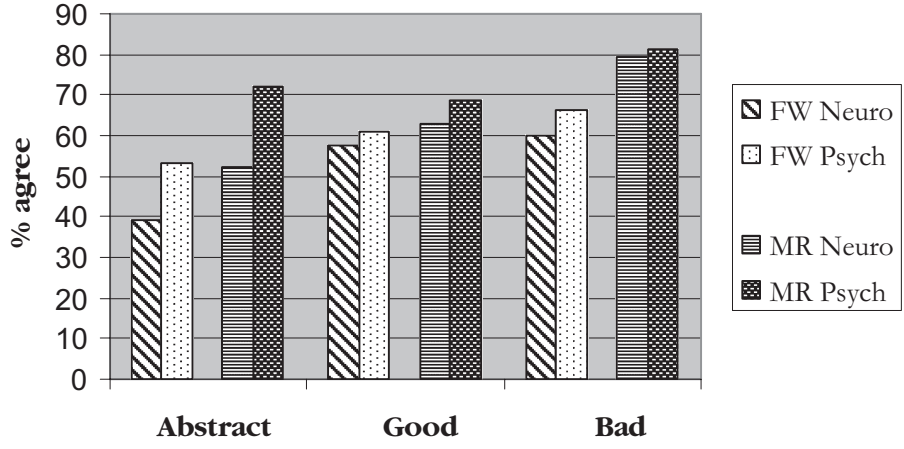

Figure 2. The effect of affect.
Percentage of participants who agreed that agents make decisions of their own free will (FW) and are morally responsible (MR) for their decisions.

In addition to comparing whether mechanism has a significant effect on judgments of FW, MR, and related concepts, we were also interested in testing whether the concreteness, or the goodness or badness, of the action being described might mediate the effect of mechanism. To this end, we ran a 2 (mechanism: Neuro, Psych) × 3 (concreteness: Abstract, Bad, Good) ANOVA for three of our five variables of interest: "up to," FW, and MR.[21] We also ran a 2 (mechanism: Neuro, Psych) × 2 (concreteness: Abstract, Bad) ANOVA for blame judgments and a 2 (mechanism: Neuro, Psych) × 2 (concreteness: Abstract, Good) ANOVA for praise judgments. All six scenarios compared in these analyses were set in the Ertan universe (we have not yet run concrete variations with Real world scenarios). For "up to," FW, and MR judgments, we found a main effect for mechanism,[22] a main effect for concreteness,[23] and no interaction effect[24] (see Figure 2). In each of the 2 × 3 comparisons, we used Tukey's Honestly Significant Difference Test to determine which conditions (Abstract, Good, and Bad) were significantly different from each other. We found that for "up to" and FW responses, mean differences were significant between the Abstract and Bad scenarios, but not between the Abstract

21. Means and SDs for: Neuro Bad (Up To: M = 3.84, SD = 1.49; FW: M = 3.98, SD = 1.55; MR: M = 4.58, SD = 1.41; Blame: 4.46, 1.41), Neuro Good (Up To: M = 3.53, SD = 1.51; FW: M = 3.59, SD = 1.64; MR: M = 3.76, SD = 1.66; Praise: M = 3.86, SD = 1.41), Psych Bad (Up To: M = 4.33, SD = 1.46; FW: M = 4.24, SD = 1.56; MR: M = 4.81, SD = 1.42; Blame: M = 4.80, SD = 1.29), and Psych Good (Up To: M = 3.98, SD = 1.50; FW: M = 3.76, SD = 1.55; MR: M = 4.11, SD = 1.49; Praise: M = 4.16, 1.26). For Neuro Ertan and Psych Ertan results, see note 12.

22. Up To: $F(1, 468) = 10.90$, $p = .001$; FW: $F(1,462) = 4.632$, $p = .032$; and MR: $F(1, 459) = 8.83$, $p = .003$.

23. Up To: $F(2, 468) = 9.482$, $p < .001$; FW: $F(2, 462) = 9.014$, $p < .001$; and MR: $F(2, 459) = 15.349$, $p < .001$.

24. Up To: $F(2, 468) = .009$, $p = .991$; FW: $F(2, 462) = .564$, $p = .569$; and MR: $F(2, 459) = 1.065$, $p = .346$.

and Good scenarios or between the Bad and Good scenarios. For MR responses, we again found a significant difference between the Abstract and Bad scenarios, but not between the Abstract and Good scenarios. Unlike "up to" and FW responses, however, the difference for MR responses between the Bad and Good scenarios was significant.[25] When comparing the results for responses to the blame question, we found a significant main effect for both mechanism[26] and concreteness,[27] but no interaction effect. For praise responses, we found no main effects or interaction effect.

Put simply, participants' responses to questions about "up to," FW, MR, praise, and blame showed a consistent pattern: lowest in the Abstract scenarios, highest in the Bad scenarios, and somewhere in between in the Good scenarios, and consistently higher in Psych scenarios than Neuro scenarios. One interpretation of these findings is that, as Nichols and Knobe have proposed, affect acts as a type of bias leading to higher judgments of FW and MR in the Bad scenarios, and to some extent in the Good scenarios, but since the Good scenarios are less affectively charged than the Bad scenarios, they induce this bias less significantly. We will suggest another interpretation below.

We will close this section by discussing a concern that anyone studying our results might have. On almost every question we asked, there was a substantial split in people's responses, with somewhere between ten to forty-nine percent of people comprising the "minority" group for each question. Such results might suggest several potentially problematic possibilities. First, it might suggest that many participants simply felt confused about the scenarios and didn't really know how to answer, leaving answers distributed in a somewhat random way across the range of options. However, for all of the main experimental questions (1–10), participants did have the option to answer "I don't know," and very few took that option (almost always less than five percent).[28]

Of course, even if participants did not *feel* confused, that does not mean they were not actually confused. Again, we did remove from analysis those who missed either of the manipulation checks. But a critic might still worry that many participants did not understand some of the scenarios or questions. This is always a possibility, and we simply suggest that further experimentation would be necessary to address it. Examination of participants' answers to the free-response questions might offer some insight. We should point out that the use of statistical analysis allows us to have greater confidence that statistically significant effects are not the result of chance factors, including people responding to questions in a haphazard or random way.

Finally, one might worry that the lack of unanimity in people's responses indicates that they simply do not have shared concepts or intuitions about these

25. In all cases that reached significance, results were significant at $p < .001$.

26. $F(1, 363) = 11.98$, $p = .001$.

27. $F(1, 363) = 26.525$, $p < .001$.

28. Indeed, the fact that a lot of participants *did* avail themselves of the "I don't know" option for some of the other experimental questions (i.e., 14, 15, 16, and 18; see Table 1) gives us some confidence that they did *not* feel pressure to avoid that answer on the main experimental questions (1–10).

issues (compare Knobe and Doris [forthcoming] and Nichols and Ulatowski [forthcoming]). Of course, it would take further research to determine if that is the best explanation for the diversity of responses, but if so, it would be an interesting discovery in its own right. It would potentially explain why philosophers bring such conflicting intuitions to debates about free will and why such debates seem to reach various impasses or "dialectical stalemates" (Fischer 1994, 83).

7. DISCUSSION

Given the history of the free will debate as well as the recent experimental work on FW and MR discussed in section 2, we formulated three main predictions to test:

(1) that most people will judge that determinism is *not* threatening to FW and MR if determinism is described in *non*mechanistic (psychological) terms.
(2) that significantly more people *will* judge determinism to be threatening to FW and MR if determinism is described in mechanistic (neuroscientific) terms.
(3) that people will significantly increase their judgments of FW and MR in response to descriptions of specific agents who perform bad acts in comparison to agents and actions described in an abstract way.[29]

We found strong evidence in support of all of these predictions. For instance, in support of prediction 1, we found that in the Psych Real scenario a significant majority of participants agreed that people would be free and responsible even if their decisions were determined by prior events, as long as those events included their thoughts, desires, and plans (see Table 1, column B). When responses are combined across the four scenarios that describe determinism in psychological terms, we see that 2/3 (sixty-six percent) of participants responded that agents make decisions of their own free will, and 4/5 (seventy-nine percent) judge that they are morally responsible for their decisions. These results closely track the results found in Nahmias et al. (2005, 2006). We suggest that as long as people are not primed to think that determinism entails mechanism, most do *not* perceive it as incompatible with FW or MR. And this suggests to us that most non-philosophers do *not* have intuitions that support Pure Incompatibilism (PI).

We hasten to add that some participants *did* express such intuitions.[30] One might argue that our presentation of determinism wasn't salient enough for many participants to recognize the threat to FW and MR that this minority recognized, or that most of our participants were unable (or unwilling) to make proper inferences that would lead them to express their underlying incompatibilist intuitions. Perhaps the participants in the *minority* were the ones who recognized the implications of determinism and drew the "proper" conclusions, while the majority failed to do so. We would be interested to see these responses supported with data

29. We did not make a specific prediction about what effect concrete *good* actions would have on people's judgments of FW or MR.

30. For instance, some in the minority explained their judgments that agents do not have MR in the Psych scenarios, with statements such as "Ertans are not responsible for their decisions because their decisions are based on events that happened before they were born."

from a systematic study (rather than suggested from the armchair). We worry, however, that getting most people to make the "proper" incompatibilist inferences may require exposing them to principles drawn from incompatibilist arguments (such as "transfer of nonresponsibility" principles), whose intuitive plausibility is contested.

We also suspect that attempts to present determinism in a way that is salient enough to significantly lower judgments of FW and MR will do so primarily by presenting determinism in such a way that it suggests other, separate threats to FW and MR. Again, we think that presenting determinism to suggest the separate threat of reductive mechanism has this effect. Here, the evidence in support of prediction 2 is relevant. When examining the combined responses to the four Neuro and four Psych scenarios, participants were significantly less likely to judge agents to be free, responsible, praiseworthy, or blameworthy in the Neuro scenarios. This effect was clearly driven by responses in the two Abstract scenarios (Real and Ertan).[31] For instance, whereas just thirty-eight percent agreed that agents have FW in the Neuro Real scenario, almost eighty-five percent agreed with the same statement in the Psych Real scenario (see Figure 1). In the Neuro Real scenarios people expressed particularly pronounced MI intuitions (see Table 1, column A; e.g., only thirty-eight percent agreed that agents deserve to be blamed for their decisions in that scenario, as opposed to seventy to eighty-six percent in the various Psych scenarios).

These results indicate that even if some people have the intuition that determinism precludes FW and MR, many *more* people have the intuition that mechanism precludes them. We think these results are best explained in terms of the psychological processes that regulate whether people engage in the mechanistic stance toward other agents. By default, humans take the participant stance toward other agents who behave in purposeful ways. When people adopt this stance toward an agent, they tend to assume that ascriptions of FW and MR are appropriate unless and until they perceive factors that are paradigmatically excusing or exempting (e.g., infancy, insanity, brainwashing, ignorance, coercion—some of these cases may themselves lead people to take the mechanistic stance toward the agent).[32] Notice, for instance, that over ninety percent of our participants agreed that normal human adults make many decisions of their own free will (see Table 1, row 17). So, when agents' decision making is described in terms of their psychological processes (e.g., thoughts, desires, and plans), people take the participant stance toward them, such that the agents are assumed to be free and responsible unless perceived to be affected by some paradigmatically excusing or exempting condition (and for most people, determinism *per se* does not serve as such as exempting condition). However, when people adopt the mechanistic stance toward an agent (for instance, when primed by a description of decision-making in terms of neural processes), then they tend to disengage from the participant stance. And

31. We will discuss below why we think these differences between Neuro and Psych scenarios diminish in the Concrete variations (Good and Bad).

32. See Strawson's (1962) discussion of excusing and exempting conditions and the "objective attitude" we take when such conditions are present.

they tend to treat the mechanistic explanations as precluding mentalistic explanations. Thus, being viewed as a mechanistic system does act as an exempting condition. As one of our participants in the Neuro Real version put it, "If the choice one makes is a reaction to the chemicals in the brain then that is not their own free will." We think that many philosophers (and others) take it that determinism is so obviously (or intuitively) incompatible with FW and MR because *purportedly* incompatibilist intuitions are instead expressions of the intuition that *mechanism* conflicts with FW and MR.[33]

We want to emphasize that the theory we are offering here, despite support from the empirical data we've collected, is still speculative (this is, after all, experimental *philosophy*). To defend the theory more fully, we would need to offer more evidence that taking the mechanistic stance does lead people to abandon the intentional stance (e.g., in our Neuro scenarios, we could ask whether the agents have beliefs, desires, plans, etc.). We would need to show that when presented with an *indeterministic* universe, people still see mechanism as a threat to freedom and responsibility. And we would need more evidence that incompatibilist intuitions get their foothold in most people (perhaps even incompatibilist philosophers!) because determinism is mis-presented in a way that leads people to take the mechanistic stance. We would also need to respond to any incompatibilist arguments that aimed to show that determinism and mechanism are *not*, as we suggest, distinct theses but that they somehow entail one another. Although we think no such argument is forthcoming, if one could be mounted successfully, then any intuitive support for the MI position would thereby support the PI position. Even in this case, however, we wonder whether the intuitive threat to FW and MR from this thesis of determinism would derive primarily from the mechanistic element of the thesis—that our mental states are not playing the right role in our actions—rather than from the deterministic element of the thesis—that there are sufficient conditions for our actions that are ultimately beyond our control.

If our theory is on the right track, it may also provide an interpretation for some of the other results we obtained. For instance, in addition to Nichols and Knobe's theory that affect influences judgments of MR, our theory might help to explain why people are more inclined to ascribe freedom and responsibility to specific agents in concrete scenarios rather than abstractly described agents. It seems likely that people more easily take the participant stance toward a specific agent performing a specific action than toward agents who are abstractly described. We found that judgments of FW, MR, "up to," praise, and blame were significantly higher in the Bad scenarios than the Abstract scenarios. And in the Good scenarios, responses were generally higher than in the Abstract scenarios (see Figure 2). It would also be useful to test this idea with scenarios describing concrete agents performing morally neutral actions (such as mowing the lawn) that are unlikely to trigger any significant affective response.[34] In the Bad scenarios, the

33. Notice that this strategy parallels historical compatibilist strategies that try to explain away incompatibilist intuitions by pointing out that determinism is being mistakenly conflated with something that is intuitively threatening to free will (e.g., coercion or external constraint).

34. All of the scenarios used in Nahmias et al. (2006) involved a specific agent performing a specific action (in some cases, good actions or neutral actions).

increased affective salience of the morally repugnant action (Smit's killing his wife) may work directly on people's judgments of MR, but it may also work by stimulating people to take up the participant stance toward the agent. In any case, the results we found for judgments of MR and blame in the Neuro Bad scenarios suggest either that participants were no longer viewing the agent (Smit) from the mechanistic stance or that, in such cases, they simply do not view the mechanistic stance as inconsistent with ascriptions of MR.[35]

We are, however, less inclined than Nichols and Knobe to treat the effect of affect on people's judgments of MR as a "performance error" that undermines rational judgments of MR. While we think that this may be the right way to interpret some people's responses to extreme cases, such as heinous murders, it would be a mistake to conclude that, in general, emotional influences on judgments about responsibility or morality should be considered performance errors. Rather, it may be that certain emotional responses should be considered *enabling factors* that engage the cognitive processes we employ from *within* the participant stance to make discriminations about a particular agent's responsibility for particular actions (and about excusing or exempting conditions). Perhaps when we make abstract judgments about agents in general or universal conditions, we sometimes employ these cognitive processes in suboptimal ways or even employ entirely different cognitive processes. Strawson's (1962) views are relevant here. He argues that the idea that there could be *universal* exempting conditions (such as the truth of determinism) for judgments of moral responsibility misunderstands the nature of such judgments, because our practices of making such judgments are appropriately attuned to differences *between* cases *within* a community of agents employing such practices. Furthermore, Strawson suggests that rather than seeing emotions as improperly biasing judgments of responsibility, it would be more appropriate to regard them as enabling factors. Our reactive attitudes, such as resentment and gratitude, are the *basis* of our judgments of responsibility; they are not what follows only *after* we have made a cold cognitive calculation of whether an agent satisfies certain metaphysical conditions for responsibility.[36]

Finally, our theory also suggests an interpretation for the result that the real world scenarios evoked more pronounced judgments in participants than the alternate world scenarios—most notably, the significantly higher judgments of FW and MR in the Psych Real scenario than the Psych Ertan scenario. It may be that participants were more likely to take the participant stance toward even abstractly described *humans* who are making decisions based on their psychological states

35. This might explain why some people are inclined to ascribe MR and blame to the mentally ill, mentally disabled, and children when they carry out heinous crimes.

36. This issue also connects up with recent debates about the cognitive mechanisms underlying moral judgments (Greene 2003). Various types of evidence suggest that people's judgments about the moral goodness or badness of events are significantly influenced by their emotional responses to such events. These emotions might be viewed as biases undermining proper moral judgments. But it seems appropriate to adopt a more fine-grained theory that tries to determine *when* emotional influences are biases and when they *enable* the relevant competencies. For instance, are we being *biased* to make irrational judgments when we consider moral issues by reading literature or watching films that evoke strong emotional responses about those issues?

than toward alien creatures (Ertans), even though they are described as being similar to humans (in their psychological processes and in other ways). At least without the special effects of a film like *E.T.*, people may simply be slow to take the participant stance toward nonhuman creatures.

8. CONCLUSIONS

Our primary purpose in this paper has been to present the results of our experiments designed to map people's intuitions about free will and moral responsibility. We also hope to have provided reasons to believe that such experimental projects can be valuable for philosophical debates about these topics. We believe that our results provide a way to explain away the common claim that most people have incompatibilist intuitions. But we fully recognize the need for further philosophical consideration and empirical research to defend this theory. If we are right, however, we suggest that the philosophical debates about free will, which have focused so much attention on the thesis of determinism, will need to focus more attention on the thesis of mechanism, the mind–body relation, and the relevance of the increasingly mechanistic understanding of human behavior offered by the sciences.

Unless the philosophical debates about the compatibility of *determinism* with FW and MR are clearly translatable into debates about the compatibility of *mechanism* with FW and MR—and we don't think they are—they risk being less relevant to people in the real world than, for instance, the burgeoning discussion of freedom and responsibility surrounding recent work in the sciences of the mind.[37] The default position taken by many of these scientists (and the media who cover them) is that by explaining how we are mechanistic systems, they are thereby "proving" that our free will is an illusion and that we are not morally responsible (see opening quotations). They reach this conclusion because they seem to take it as obvious (intuitive) that mechanistic systems cannot have FW or MR. We worry that these views, which are problematic in several ways, are making their way into our cultural milieu, and they have the potential to influence people's beliefs and practices in more significant ways than more philosophically informed discussions of these issues.

One reason that this trend is problematic is that we are not convinced that the MI intuition is *true*. Despite people's intuition that taking the mechanistic stance toward a system (e.g., a human being) precludes taking the intentional or participant stances toward the same system, it's not clear that the same system cannot be properly understood as both mechanistic and mentalistic (or intentional).[38] If we are right about this, then understanding the way people tend to think about these issues, and the psychological processes that lead them to think this way, will be crucial for revising their thinking so that they may come to accept the fact—if it is a fact—that we can be both mechanistic systems *and* mentalistic

37. A striking example of this point is the current trend in legal defenses to build their arguments around neuroscientific evidence. As far as we know, only Clarence Darrow has attempted to mount a defense based on the grounds that if the universe is deterministic, then a client should not be held morally responsible.

38. See, for example, Dennett (1973).

systems with at least the possibility of having genuine FW and MR. It may be that the intuitions that support MI are deep-seated, like the intuitions that supported the conclusion that the earth is the stable center of the universe. And just as it required Galileo's explanation of how the earth, despite appearances, could possibly be moving in order to make the Copernican conclusion palatable, it might require another Copernican revolution to explain how mechanistic systems can possibly be conscious, intentional, rational systems. But the arguments for this "Mechanism Compatibilism" will have to go beyond standard compatibilist arguments, since those arguments *begin* with the assumption of agents who are conscious, intentional, and rational and then try to show that such agents can be free and responsible in a deterministic world.

We are thus inclined to suggest revisionism about free will *via* a revision of ordinary people's understanding of the nature of the mind, consciousness, and rationality. It would take convincing explanations for why the relevant mental states can exist (and be causally efficacious) even if they are also best understood as consisting of physical, mechanistic (e.g., neural) processes. But if these explanations did convince people to revise their understanding of the mind, we do not think it would take yet *another* revision of their conceptions of freedom and responsibility if we then discovered that determinism is true. That is, if the *neuroscientists* tell us that our behavior is caused by our brains, people will need to be convinced that it is nonetheless possible for our behavior to be caused by *us* (i.e., our mental states). But if the *physicists* then tell us that our universe is ultimately deterministic, we will not then need to convince people that we are free and responsible agents. Philosophers engaged in the free will debate should hence consider whether the problems posed by the possibility of mechanism are more pressing than the potential problems posed by the possibility of determinism.

APPENDIX: FURTHER QUESTIONS AND ANSWERS

Because we used an online survey methodology, we were able to obtain substantially more data than we have presented in the body of this paper. And because much of this data offers interesting information about several philosophical issues, we have included this appendix, in which we will discuss some other questions that our data address and some of the other results we obtained. We have presented these results in summary form in Table 1.[39]

(1) *Are there interesting relationships between people's responses to the various ways we asked questions about agency?*

Yes, people's responses to questions about whether agents make decisions of their own *free will* (row 2) and people's responses to questions about whether agents' decisions are *up to* them (row 1) tracked each other very closely across all eight scenarios (generally rates of agreement were within five percent, with the

39. Note that we have not yet examined another large data set, the open-ended explanations participants offered about their responses to the questions about free will, moral responsibility, and meaningful lives.

totals across all scenarios within two percent). To confirm this relationship, we ran Wilcoxon Signed Rank Tests, which allow within-subject comparisons between responses to questions. These statistical analyses confirmed that there were no significant differences between participants' responses to these questions about FW and "up to" in *any* of the scenarios. One plausible interpretation of these results is that people are interpreting these two questions to mean roughly the same thing. If that claim is right, then future surveys of people's intuitions about free will may be able to use the very ordinary phrase "up to" to replace the more technical and contentious term "free will."[40] However, it would require further testing to rule out the possibility that, in response to certain scenarios, people would say that an agent's decision *was* up to her even though she did *not* make it of her own free will (or, less plausibly, she made a decision of her own free will, but her decision was *not* up to her). Scenarios could be developed to test this possibility.

People's responses to the question of whether agents *control* what they do (row 6) also tracked responses to free will relatively closely, though not quite as closely as "up to" did, and we have less data on the "control" question, since we only used it in four scenarios. We are also inclined to think that people's conception of control would *not* track their conception of free will in certain cases (e.g., people would likely say that a thermostat controls a room's temperature but not of its own free will).

(2) *Did people's responses to the question of whether agents are morally responsible for their decisions track their responses to questions about free will (and "up to")?*

No. Across all scenarios, judgments about MR were consistently higher than judgments about FW (and "up to"). Levels of agreement to statements about agents' MR (row 3) are, in most cases, significantly higher than levels of agreement to statements about agents' FW (row 2). In some cases, the difference is small, but in other cases, it jumps to nearly twenty percent (e.g., in Psych Ertan, Neuro Bad, and Psych Bad). Summed across all scenarios, judgments of MR are twelve percent higher than judgments of FW. Statistical analyses show that there is a significant difference between mean judgments of FW and MR in all scenarios except the Good scenarios and also between "up to" and MR judgments in all scenarios except the Good scenarios and Neuro Real (and in the exceptional cases, the trend is still in the same direction—i.e., MR higher than FW and "up to"). Notice also that judgments about whether agents deserve praise or deserve blame for their decisions generally tracked judgments about MR (there were no statistical differences in 9 of 12 comparisons).[41]

40. Nichols and Knobe (forthcoming) avoided asking people about free will precisely because they thought it was a technical notion. Philosophers often discuss free will in terms of the idea that our actions or decisions are "up to" us.

41. We also compared judgments of blame to judgments of praise in the abstract scenarios (we did not ask both questions in the concrete scenarios). While there were no significant differences between praise and blame in the Psych cases, in both Neuro Real and Neuro Ertan, praise judgments were significantly higher than blame judgments. This is consistent with some literature

We suspect that these higher judgments about MR compared to FW may be due in part to the ambiguous nature of concepts of moral responsibility and desert. When people say that someone is morally responsible for his action or that he deserves blame, they may mean, or wish to suggest, that he has done something bad and needs to be blamed and punished to prevent him (and others) from repeating such actions. They may also mean that he has freely done something blameworthy such that he deserves to be punished for what he has done, regardless of any future effects it might have. But these forward-looking (deterrent and prevention) conceptions and backwards-looking (retributive) conceptions of responsibility and desert may come apart, such that different people respond to relevant questions in different ways depending on which conception they are using or which is most salient to them at the time.

For instance, it is possible that in the Abstract scenarios, the backwards-looking (retributive) conception is more salient because there is no specific agent who has performed an action in such a way that he needs to be held responsible for it in order to prevent (or to encourage) him and others from performing such actions. However, in the Bad and Good scenarios, the forward-looking conception of responsibility may be more salient, because there is a specific agent (Smit) who has performed a bad action that needs to be deterred or who has performed a good action that needs to be encouraged. This might help to explain why judgments of MR (and blame) tend to go up in the Bad scenarios compared to the Abstract scenarios (across both Neuro and Psych cases).

The concepts of moral responsibility and desert are foundational to the philosophical debates about free will (e.g., most philosophers analyze free will in terms of the powers required to be held responsible for one's actions such that one deserves praise or blame for them). As such, it would be helpful to understand more fully how ordinary people understand these concepts and how our practices employ them.[42]

It is in part because the concepts of moral responsibility and desert are potentially ambiguous in these ways that we asked questions 14 and 15 about punishment. These questions were designed to probe whether people think criminal punishment should have, in addition to its deterrent functions, a retributive function (to make criminals suffer because they deserve it for what they have done), and whether the truth of determinism (as described in the scenarios) would suggest that the retributive function would be inappropriate.

in psychology (e.g., Malle 2006), which suggests judgments of praise are less sensitive to excuses and exemptions than judgments of blame, and also with Wolf's (1990) "asymmetry thesis," which suggests that judgments of blame are sensitive to the agent's alternatives in a way judgments of praise are not.

42. Another foundational issue in the debate involves modal concepts (such as ability). To explore this issue, we asked questions 12 and 13 about whether agents in the scenario *might* (or *can*) make more than one decision, once certain mental (or neural) states occur. Though answers varied across situations, in six of the scenarios, about seventy percent of participants answered "yes" to this question (see Table 1, rows 12 and 13). It would be helpful to explore this issue more fully. However, these concepts are difficult enough for philosophers to analyze; we have found it very difficult to frame them in useful ways for non-philosophers .

(3) *How did participants respond to these questions about retributive punishment?*

More people answered "I don't know" to these questions than almost any other, generally twenty to thirty percent (see Table 1, rows 14–15). That indicates either a problem with the questions or a genuine puzzlement in people's beliefs about this complex issue, or (as we suspect) both. We were somewhat surprised that relatively few people indicated that they think criminals deserve to suffer—rarely more than fifty percent—but it may be that the word "suffer" is problematic here (we cannot think of a viable substitute to reveal the relevant intuitions). Across all scenarios, there was a trend towards lower agreement with retributive punishment in the mechanistic scenarios (especially Neuro Ertan), and it would be interesting to examine whether that trend was driven (as we predict) by participants who judged that agents in the scenario did not have free will, the idea being that retributive punishment is inappropriate for agents without FW (perhaps even if they are considered morally responsible and deserving of blame in some sense of those concepts).

(4) *Do people think you need a soul to have free will?*

An interesting question about the psychological mechanisms underpinning the mechanistic stance and the mentalistic (intentional and participant) stances is whether most people think that having mental states (or at least having FW) requires having a non-physical soul (see Bloom 2004). Presumably, interesting differences regarding these issues would emerge between people with different cultural and religious backgrounds. With these issues in mind, we included questions about religious affiliation and religiosity and questions 16 and 18 regarding their beliefs about souls. We have not yet examined whether there are interesting correlations between people's responses to these questions and their responses to the other questions about FW and MR. The majority of people indicated that they believed humans have non-physical souls (row 16) and the vast majority responded that humans have free will (row 17). Question 18 was added for four scenarios to examine whether people believe we have free will *only* because we have souls. Though we suspect it was difficult for participants to interpret this question, since so many answered "I don't know," we were surprised that so few (twenty percent) indicated that we have free will *only* because we have non-physical souls. We hope to explore these questions in future research.

(5) *Did people think that determinism would undermine the possibility for personal relationships, meaningful lives, or ethical behavior?*

We asked questions 8–10 about these issues because some philosophers have suggested that people believe determinism conflicts with free will and that losing our belief in libertarian free will would radically change the way we think about meaning, morality, and love (see, e.g., Smilansky 2000). We found, however, that a significant majority of people responded that agents in a deterministic universe *could* have meaningful lives and deep personal relationships (except in the Neuro Real scenario; see rows 8–10; responses to the question about acting ethically were

less revealing). This result provides some support for the idea that people do not take determinism *per se* to be a threat to the possibility of having meaningful lives or deep relationships.[43]

On the other hand, we also found that people's responses to questions about FW and MR correlated significantly with their responses to the questions about meaningful lives and relationships. That is, participants who disagreed with the statements about FW and MR indicated less agreement with the statements about meaningful lives, relationships, and acting ethically. This suggests some support for Smilansky's view that losing our belief in free will might have adverse effects on our conception of our lives and relationships—that they would lose a certain sort of depth and meaning for us. However, it is not clear that people have a *libertarian* conception of free will in mind when they make these judgments. Finally, the results also indicate that many participants were willing to say both that people (or Ertans) could have meaningful lives and relationships while also saying that they do *not* have FW or MR.[44]

This brings us back to our initial discussion of the relevance of ordinary intuitions and practices to the philosophical debates. We need to explore more fully how people's conceptions of themselves and others as free and responsible agents impact the things they care about most deeply. If the philosophical debates are contesting concepts that are *not* connected to what people care about, then those debates risk becoming merely academic exercises.[45]

REFERENCES

Birnbaum, Michael, ed. 2000. *Psychological Experiments on the Internet*. New York: Academic Press.

Bloom, Paul. 2004. *Descartes' Baby*. New York: Basic Books.

Dennett, Daniel. 1973. "Mechanism and Responsibility." In *Free Will*, ed. Gary Watson, 150–73. New York: Oxford University Press. 1982.

———. 1987. "True Believers: The Intentional Strategy and Why It Works." In *The Intentional Stance*, ed. Daniel Dennett, 13–36. Cambridge, MA: MIT Press.

Ekstrom, Laura. 2002. "Libertarianism and Frankfurt-Style Cases." In *The Oxford Handbook of Free Will*, ed. Robert Kane, 309–22. New York: Oxford University Press.

Fischer, John Martin 1994. *The Metaphysics of Free Will*. Oxford: Blackwell Publishers.

Greene, Joshua. 2003. "From Neural 'Is' to Moral 'Ought': What Are the Moral Implications of Neuroscientific Moral Psychology?" *Nature Reviews Neuroscience* 4: 846–49.

Kane, Robert. 1999. "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism." *Journal of Philosophy* 96: 217–40.

Kim, Jaegwon. 1998. *Mind in a Physical World*. Cambridge, MA: MIT Press.

Knobe, Joshua, and Doris, John. Forthcoming. "Strawsonian Variations: Folk Morality and the Search for a Unified Theory." In *The Handbook of Moral Psychology*, ed. John Doris. Oxford: Oxford University Press.

43. We wonder whether the fact that people's answers to these two questions tracked each other very closely—a within-subject analysis confirms this—is reason to believe that most people believe that one can have a meaningful life if and only if one can have deep personal relationships!

44. For more discussion of these issues, see Nadelhoffer and Feltz (2007).

45. Nahmias is primary author; Coates and Kvaran contributed equally. For helpful comments, we would like to thank Andrea Scarantino, Bradley Thomas, and Jason Turner. We are indebted to Erin McClure for her advice regarding statistical analyses.

Lewis, David. 1986. *The Plurality of Worlds*. Oxford: Blackwell Publishers.
Malcolm, Norman. 1968. "The Conceivability of Mechanism." In *Free Will*, ed. Gary Watson, 127–49. New York: Oxford University Press. 1982.
Malle, Bertram. 2006. "The Relation between Judgments of Intentionality and Morality." *Journal of Cognition and Culture* 6: 61–86.
Monterosso, John, Royzman, Edward, and Schwartz, Barry. 2005. "Explaining Away Responsibility: Effects of Scientific Explanation on Perceived Culpability." *Ethics and Behavior* 15(2): 139–58.
Nadelhoffer, Thomas, and Feltz, Adam. 2007. "Folk Intuitions, Slippery Slopes, and Necessary Fictions: An Essay on Saul Smilansky's Free Will Illusionism." *Midwest Studies in Philosophy* 31: 202–13.
Nadelhoffer, Thomas, and Nahmias, Eddy. 2007. "The Past and Future of Experimental Philosophy." *Philosophical Explorations* 10(2): 123–49.
Nahmias, Eddy. 2006. "Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism." *Journal of Cognition and Culture* 6(1–2): 215–38.
Nahmias, Eddy, Morris, Stephen, Nadelhoffer, Thomas, and Turner, Jason. 2005. "Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility." *Philosophical Psychology* 18: 561–84.
———. 2006. "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research* 73: 28–53.
Nichols, Shaun, and Knobe, Joshua. Forthcoming. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Nous*.
Nichols, Shaun, and Roskies, Adina. "Bringing Moral Responsibility Down to Earth" (unpublished).
Nichols, Shaun, and Ulatowski, Joe. Forthcoming. "Intuitions and Individual Differences: The Knobe Effect Revisited." *Mind and Language*.
Smilansky, Saul. 2000. *Free Will and Illusion*. New York: Oxford University Press.
Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Clarendon.
Strawson, Peter. 1962. "Freedom and Resentment." In *Free Will*, ed. Gary Watson, 59–80. New York: Oxford University Press. 1982.
Swinburne, Richard. 1989. *Responsibility and Atonement*. Oxford: Clarendon Press.
Turner, Jason, and Nahmias, Eddy. 2006. "Are the Folk Agent-Causationists?" *Mind and Language* 21(5): 597–609.
van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.
Vargas, Manuel. 2005. "The Revisionist's Guide to Moral Responsibility." *Philosophical Studies* 125(3): 399–429.
Wolf, Susan. 1990. *Freedom within Reason*. New York: Oxford University Press.