CHAPTER 29

INTUITIONS ABOUT FREE WILL, DETERMINISM, AND BYPASSING

EDDY NAHMIAS

It is often called "the problem of free will and determinism," as if the only thing that might challenge free will is determinism, and as if determinism is obviously a problem.[1] The traditional debates about free will have proceeded accordingly. Typically, incompatibilists about free will and determinism suggest that their position is intuitive or commonsensical, such that compatibilists have the burden of showing how, despite appearances, the problem of determinism is not really a problem. Compatibilists, in turn, tend to proceed as if showing that determinism is not a problem thereby shows that we have free will, as if determinism is the only thing that might threaten free will.

Robert Kane captures both of these elements of the traditional debate, first suggesting that the truth of compatibilism would demonstrate that we have free will: "If compatibilists are right, we can have both freedom and determinism, and need not worry that future science will somehow undermine our ordinary conception that we are free and responsible agents"; and then suggesting that compatibilism is highly counterintuitive:

> In my experience, most persons resist the idea that free will and determinism might be compatible when they first encounter it. The idea that determinism might be compatible with free will looks at first like a "quagmire of evasion," as William James called it, or a "wretched subterfuge" as Kant called the compatibilism of Hobbes and Hume. If compatibilism is to be taken seriously by ordinary persons, they have to be talked out of this natural belief in the incompatibility of free will and determinism by means of philosophical arguments. (Kane 2005a, 12–13)

In this chapter, I reject both of these elements of the traditional debate; the question of whether we have free will should neither begin nor end with the so-called problem of determinism. I present and discuss evidence from a variety of studies that suggests that incompatibilism is not particularly intuitive. Most people do not have to be talked out of incompatibilism but rather talked *into* it. This provides some reasons—though certainly not decisive reasons—to think that compatibilism is true. I conclude by pointing out that, even if compatibilism were true, it would not dissolve the problem of free will, because there are problems other than determinism that need to be confronted—namely, challenges to free will suggested by discoveries in neuroscience and psychology. The threats to free will suggested by these sciences are distinct from the traditional threat of determinism, and *they* are the ones that "ordinary persons" find intuitively threatening to free will. In fact, I will argue that the reason incompatibilism about free will and determinism *appears* to be intuitive is that determinism is often and easily *mis*understood to involve these distinct threats to free will—threats that suggest that our rational, conscious mental activity is *bypassed* in the process of our making decisions and coming to act.

Whose Intuitions Matter and Why?

All this talk of what is intuitive should make one wonder: "intuitive to whom?" Above, Kane mentions "ordinary persons" as they first encounter the problem of free will and determinism. Timothy O'Connor (2000, 4) likewise writes, "Does freedom of choice have this implication [that determinism must be false]? It seems so to the typical undergraduate on first encountering the question." And Galen Strawson (1986, 89) claims that it is "in our nature to take determinism to pose a serious problem for our notions of responsibility and freedom." Many other philosophers agree with these claims that ordinary people have pretheoretic, commonsensical intuitions that determinism is incompatible with free will, though some compatibilists argue that it is *their* position that is intuitive.[2] Many philosophers on all sides of the debate seem to think that nonphilosophers' understanding of free will is relevant to the philosophical debate. There are at least two good reasons why they *should* think this.

First, philosophers remain deadlocked about how to define free will and whether the proper conception of free will is one which entails that it is, or is not, compatible with determinism. These "dialectical stalemates" (Fischer 1994) also permeate their views about the premises, principles, and crucial thought experiments used in opposing arguments, such as Frankfurt cases, interpretations of the ability to do otherwise, transfer principles (e.g., Beta), interpretations of the fixity of the past and laws, and manipulation arguments. In the face of these stalemates, it is not surprising that each side has suggested that its views are supported by commonsense intuitions. Such claims allow philosophers to motivate their own position and to shift the burden of proof onto their opponents, unless of course one takes the view that such appeals to folk intuitions are simply irrelevant or inappropriate.

However, these appeals to pretheoretic intuitions are *not* irrelevant or inappropriate. The second reason philosophers should be interested in ordinary intuitions is that the target concept of interest to most philosophers is the type(s) of freedom or control relevant to holding oneself and others morally responsible for their actions—that is, whether we have the sort of free will required to *deserve* praise and blame, reward and punishment. This sort of free will is generally associated with other things people care about, such as autonomy, self-development, creativity, morality, meaningful lives, and human relationships (see, e.g., Kane 1996, ch. 6). In Daniel Dennett's (1984) terms, philosophers should be interested in "the varieties of free will worth wanting." If we are interested in understanding the concept of free will that is intimately connected with these significant concerns of nonphilosophers, and if we assume that important features of this concept can be illuminated by ordinary usage and intuitions about relevant cases, then we should be interested in understanding the relevant usage and intuitions of nonphilosophers.

If debates about free will are not to focus on a technical concept, about which only the "expert" intuitions of trained philosophers are relevant, then the question turns to how we can best garner information about the relevant intuitions of nonphilosophers. One possibility is for philosophers to assume that their *own* intuitions about the relevant cases (including thought experiments) and concepts are representative of ordinary intuitions. We should worry, however, about whether philosophers, especially those "embedded" in the free-will debate, are likely to make claims about what is intuitive—and to create and react to thought experiments—in ways that reflect their own theoretical commitments. Perhaps because they recognize this worry, philosophers often allude to the intuitions of their students and informal polls of them. Here the worry is that students will be influenced by the way their teachers present the issues or their responses will be interpreted in ways that support their teachers' own theories. The way people consider the potential problem of determinism "when they first encounter it" may depend largely on how this first encounter is orchestrated.[3]

Assuming, then, that ordinary intuitions about free will are relevant to the philosophical debates, we should use more systematic and controlled approaches to elucidating those intuitions, such as the ones employed by the emerging field of "experimental philosophy." Broadly speaking, experimental philosophers first use empirical methodologies to obtain information about the way people think about philosophical problems, and then they consider how such information sheds light on philosophical debates (see Nadelhoffer and Nahmias 2007; Knobe and Nichols 2008). Currently, the most common method employs surveys of nonphilosophers that present them with various scenarios and questions, though other methods can be and have been deployed. Experimental philosophers do not claim that a richer understanding of what nonphilosophers actually think can alone resolve philosophical debates. But such empirical information can be very useful for (at least) three purposes in debates about free will and moral responsibility.

First, it can serve to correct assumptions that philosophers have made about what is intuitive to nonphilosophers. For instance, the assumption has usually been that compatibilism is counterintuitive, a "quagmire of evasion," and hence the burden of proof is on compatibilists. In this context, evidence showing that incompatibilism is *not* pretheoretically intuitive to most people would help shift the burden of proof back onto incompatibilists. I think this shift would be particularly significant because incompatibilist theories employ a conception of free will that is more "metaphysically demanding" than compatibilist theories, requiring, at a minimum, indeterminism in the right time and place and, perhaps, agent-causal powers. If ordinary intuitions do not support this libertarian theory of free will or the premises and principles that motivate it, then it is unclear why it should be adopted, given its metaphysical demands. Incompatibilists could, of course, argue that their theory of free will requires a revision of ordinary thinking, but then they should make explicit the theoretical advantages their view offers that offset the costs of such revision (see Nahmias, Morris, Nadelhoffer, and Turner 2006).

Indeed, a second reason that accurate information about ordinary intuitions is valuable is to clarify when, and to what extent, competing philosophical theories are systematizing our relevant beliefs, concepts, and intuitions or, conversely, when they are suggesting revision of them (see Vargas 2009). Incompatibilists often suggest that compatibilist theories are revisionist, and compatibilists sometimes agree. But we cannot know whether this is true without more complete information about the folk concept or theory that is supposedly being revised. Furthermore, because people's relevant beliefs may vary both within and between cultures, the revisionist project may be more complicated than sometimes suggested. Here, it will also be useful to examine the way people's intuitions about freedom and responsibility correspond with their psychological traits, religious beliefs, cultural setting, and various other factors (e.g., gender, age, or education).

Third, we can also look at the way people's responses differ depending on various controlled differences among cases. Doing so allows experimental philosophers to obtain evidence that sheds light on the psychological sources of people's intuitions about philosophical issues. For instance, determinism can be presented in subtly different ways to see what effects that has on people's judgments about free will and moral responsibility. Most of the experiments I describe below employ this methodology. The results suggest interesting conclusions about which factors lead people to see agents as lacking free will and moral responsibility. I take the evidence to show that most nonphilosophers do *not* take determinism, properly understood, to threaten free will. At the same time, I think the evidence helps to explain why it *seems* that people have incompatibilist intuitions when in fact they do not.

Table 1 Summary of Results from Nahmias, Morris, Nadelhoffer, and Turner (2006)

Subjects' judgments that the agent…	**Scenario 1**	**Scenario 2**	**Scenario 3**
	(Jeremy)	(Fred & Barney)	(Jill)
…acts of own **free will**w	76% (robbing bank)	76% (stealing)	66%
	68% (saving child)		
	79% (going jogging)	76% (returning)	
…is **morally responsible**	83% (robbing bank)	60% (stealing)	77%
for action	88% (saving child)	64% (returning)	

Is Incompatibilism Intuitive?

The initial studies I carried out with Stephen Morris, Thomas Nadelhoffer, and Jason Turner (2005, 2006) were designed primarily with the first goal in mind, to test the assumptions philosophers had made about ordinary people's intuitions.[4] We wanted empirical evidence about whether it is true that "we come to the table, nearly all of us, as pretheoretic incompatibilists" (Ekstrom 2002, 310). We designed three different scenarios describing determinism and asked participants about agents' free will and responsibility in them. One scenario involved a Laplacean supercomputer that could use information about the state of the universe and the laws of nature to perfectly predict the future—to "look at everything about the way the world is and predict everything about how it will be with 100 percent accuracy." Despite the fact that the computer accurately predicted *before* the birth of the agent, Jeremy, what he would do at a particular future time, a statistically significant majority of participants responded that Jeremy acted of his own free will and was morally responsible for his action, whether his action was bad (robbing a bank) or good (saving a child) (see Table 1 below).[5]

In another scenario, we presented determinism in terms of "one's genes and environment completely caus[ing] one's beliefs and values," such that if two identical twins, Fred and Barney, one of whom was adopted by a family that raised him to be selfish, and the other to be generous, had instead been raised by the other's family, each would have ended up with the other's beliefs and values and would have been caused to behave accordingly (in the scenario, either stealing or returning a found wallet). Most people said that Fred and Barney acted of their own free will and were, respectively, morally blameworthy or praiseworthy.

Our third presentation described determinism in terms of a "rollback" universe:

> Imagine there is a universe (Universe C) that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of

> nature cause the exact same events for the entire history of the universe, so that every single time the universe is re-created, everything must happen the exact same way. For instance, in this universe a person named Jill decides to steal a necklace at a particular time and then steals it, and *every* time the universe is re-created, Jill decides to steal the necklace at that time and then steals it.[6]

Again, a statistically significant majority of participants responded that, despite the deterministic nature of the universe, Jill has free will and is morally responsible. My interpretation of these results is that most nonphilosophers do *not* have what Derk Pereboom (2001, 89) calls "the incompatibilist intuition" or what Kane (1999a, 217) calls "natural incompatibilist instincts." Of course, it might be that these purported incompatibilist intuitions could be "uncovered" by explaining to participants exactly why determinism is supposed to be a problem for free will. Perhaps most people need to be shown what the minority in our surveys may already see—for instance, that determinism is inconsistent with the requisite ability to do otherwise or that determinism entails that there are sufficient conditions for one's decisions that exist before one is born. Of course, philosophers disagree about whether determinism is inconsistent with the sort of ability to do otherwise that is *relevant to free will* and about whether free will and responsibility require that there *not* be sufficient conditions for one's decisions that one is not responsible for (e.g., whether "transfer of nonresponsibility" principles such as Beta are valid), and they disagree about the intuitiveness of these issues. Nonetheless, one response to these results is that most participants simply fail to understand the deterministic element in the scenarios or fail to draw implications from determinism that they would accept with some guidance.[7]

On the other hand, it might be that those participants in the majority, who respond that the agents in the scenarios *are* free and responsible, are expressing an intuition compatibilists often emphasize—namely, that even if our decisions and actions are part of a deterministic chain of events, our deliberations and decisions are a crucial *part of* that chain of events and can still play the right sort of causal role in our actions for us to be responsible for those actions. Most participants probably do not explicitly think in these terms when they are interpreting the scenarios; rather, they may implicitly recognize that there is nothing in the scenario that rules out the relevance of the agent's psychological processes playing the relevant role in the agent's decisions and actions.

It may also be that many of those participants in the minority, who *do* say that the agents lack free will and responsibility, do so because they assume that the scenario precludes a proper causal role for agent's beliefs, desires, and decisions. That is, they may assume that determinism entails that psychological processes—of the sort that compatibilists highlight as essential to free and responsible agency—are *bypassed* by other causal processes. For instance, they may assume that the fact that the Laplacean computer can predict everything based on the current state of the universe and the laws of nature means that everything that happens is completely caused by physical events, which the computer can compute, leaving no causal work for Jeremy's psychological states, such as his conscious deliberation. Perhaps some people assume that Jeremy will rob the bank no matter what he wants, thinks, or tries to do. When people read that Fred and Barney exist in a world where everything they think and do is "caused completely by the combination of one's genes and one's environment," perhaps some interpret this to mean that, because these features external to the agent's psychology are sufficient for the agent's actions, the agent's psychology is irrelevant to what he does.[8]

Thinking in terms of this sort of bypassing of the agent's relevant mental life may be what drives many of the minority incompatibilist responses. It may be hard to imagine that people are reading such bypassing of the agent's mental activity into the scenarios, especially when the scenarios use the language of "decisions," "beliefs and values," and "actions." However, as shown below, depending on how one describes determinism, people can be surprisingly disposed to interpret it to involve such bypassing.

An Error Theory for Apparent Incompatibilist Intuitions

Whether an agent is free and responsible depends crucially on whether her decisions and actions are the result of processes that go "through" her, rather than "around" her, and more specifically, whether they go through, or around, the relevant processes within her. In general, an agent's mental states and events—her beliefs, desires, or decisions—are bypassed when the agent's actions are caused in such a way that her mental states do not make a difference to what she ends up doing. One way to understand this idea is that the agent would end up doing what she does regardless of what she had thought or wanted or decided.[9]

The idea of bypassing can be made more specific by focusing on the particular psychological capacities that compatibilists highlight, such as responsiveness to reasons (Fischer 1994), higher-order volitions (Frankfurt 1971), or reflective self-governance (Scanlon 1998). But if general psychological states and processes, such as beliefs, desires, and conscious deliberations, are bypassed, then more specific compatibilist capacities will be bypassed as well. Free will can of course be compromised even if one's psychology is not entirely bypassed; for instance, if one is hypnotized, brainwashed, or deceived to believe, desire, or deliberate in certain ways, or perhaps if one acts on compulsive or addictive desires. So, lack of bypassing is not sufficient for free and responsible agency, though it is necessary. At least, lack of bypassing is necessary on most philosophical accounts of free will, both compatibilist and libertarian. Discussing one version of bypassing, Kane (2005b) points out: "If conscious willing is illusory or epiphenomenalism is true, *all* accounts of free will go down, *compatibilist and incompatibilist*."

Bypassing also seems to be a particularly intuitive threat to free will and responsibility. It seems to drive the intuition that one is not free or responsible if one is coerced or constrained, in which case one has to do something even though one does not want to do it or believes that one should not do it. Bypassing of one's rational capacities seems to drive the intuition that certain types of insanity, compulsion, manipulation, and indoctrination undermine (or at least mitigate) one's freedom and responsibility. The compulsive hand washer decides that today she will only wash her hands twice, yet finds that decision having no influence on her behavior. The person who escapes the cult wonders how he could have believed such crazy things, how his rational thinking could have been blinded by the cult leader and group effects. Bypassing may also be the worry behind fatalism, interpreted as the view that certain things will happen no matter what one decides or tries to do. Oedipus will end up sleeping with his mother no matter how he might try to avoid this fate.

My suspicion is that what drives many people who see determinism as a threat to free will and moral responsibility is (a) their strong intuition that bypassing precludes freedom and responsibility and (b) their interpretation of determinism as entailing bypassing. If this suspicion is corroborated, then it would provide an "error theory" to explain away apparent incompatibilist intuitions, because (b) is mistaken and (a) does not support incompatibilism. Determinism is the thesis that a complete description of the past state of the universe and the laws of nature logically entails a complete description of all later states of the universe, or (not equivalently) that everything that happens has sufficient prior causes. Determinism does *not* entail, nor should it be taken to mean, that a person's beliefs, desires, and decisions make no difference to what happens, or that certain things will happen even if the past had been different, or regardless of what one tried to do. On the contrary, determinism suggests that what happens later depends on what happens earlier; which actions actually occur depends on which beliefs, desires, and decisions actually occur, at least assuming that beliefs, desires, and decisions are not causally epiphenomenal. Determinism alone does not entail epiphenomenalism (the causal irrelevance of mental states and processes) nor does it entail fatalism (interpreted as the view that certain events must happen even if earlier events—e.g., one's decisions—had been different).

So, perhaps as Kane says, "most persons resist the idea that free will and determinism might be compatible when they first encounter it," but only because they think something like this:

Description of determinism → Bypassing → No free will or responsibility

If so, then these people are not "natural incompatibilists" and their intuitive response to determinism should not be taken as support for incompatibilism. Is there evidence that lots of the people who interpret determinism as threatening free will are in fact reasoning in this way? I have run three sets of experiments that suggest that the answer is "yes."

Studies on Bypassing

Mechanism as Bypassing

With Justin Coates and Trevor Kvaran, I explored the hypothesis that people would react to a description of determinism differently depending on whether or not it was described in a way that suggests mechanistic reductionism—the view that higher-level processes (e.g., human decision making) can be completely understood in terms of lower-level mechanisms, such as neural processes. We presented participants with a description of determinism that emphasized either mechanistic causation or psychological causation (the only variations in the following version of the scenarios are 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 predicted that people would interpret the scenario involving neurobiological mechanism to suggest bypassing, presumably because they implicitly assume that a complete explanation of our decisions and actions in terms of processes in the brain would leave no causal work for agents' mental processes to do (see note 8, above). Indeed, across four different variations of the scenario, we consistently found statistically significant differences between responses to the mechanistic scenario and the psychological scenario, including participants' responses to questions about whether, assuming the neuroscientists/psychologists are correct, people have free will, people's decisions are "up to" them, people should be held morally responsible, and people deserve praise and blame for what they do. For instance, in the "real world" version presented above, we found that the vast majority of participants

Table 2 Summary of Results from Nahmias, Coates, and Kvaran (2007)

Subjects' judgments that…	1. **Psych Abstract (Real World)**	2. **Neuro Abstract (Real World)**	3. **Neuro Abstract (Alt. World)**	4. **Neuro Concrete (Alt. World)**
…agents act of their own **free will**	83%	38%	39%	60% (bad) 57% (good)
…agents' decisions are **up to** them	86%	34%	40%	62% (bad) 54% (good)
…agents are **morally responsible** for their action	89%	41%	52%	79% (bad) 63% (good)
…agents **deserve blame (praise)** for their actions	86%	38%	50%	74% (blame) 71% (praise)

who read the scenario with psychological language judged that people have free will, are morally responsible, and deserve blame. On the other hand, most participants who read the scenario with neurobiological language judged that people do not have free will, are not responsible, and do not deserve blame (see Table 2, columns 1–2). Furthermore, consistent with the results described earlier, across all versions of the nonreductionistic scenarios, the majority of participants did not express incompatibilist intuitions—namely, most responded that agents in deterministic worlds could have free will, be morally responsible, and deserve praise and blame.[10]

Another intriguing finding in these studies is that participants' judgments of free will and responsibility were consistently higher in scenarios that describe specific (named) agents performing specific actions than in *abstract* scenarios, like the version above, that describe agents in general. The *concrete* versions add details to the final paragraph about an agent who decides to kill his wife so that he can marry his lover (or in a positive version, decides to donate a large sum of money to an orphanage in his community). In most cases, especially the mechanistic ones, judgments of the agent's free will, responsibility, praise, and blame were significantly higher in response to the concrete scenarios than the abstract scenarios. For instance, in an alternative universe version of the abstract mechanistic scenario (see Table 2, column 3), most participants responded that agents do not act of their own free will and half said they do not deserve to be blamed, but in concrete versions of this mechanistic scenario (column 4), significantly more responded that the agent acted of his own free will and that he deserves blame (in the negative scenario) or praise (in the positive scenario). One way to interpret these differences is that the concrete cases, at least the ones with negative actions, tend to bias people in such a way that they neglect the deterministic (and mechanistic) features of the scenario (see discussion of Nichols and Knobe [2007], below). Another interpretation is that the mention of specific persons making decisions and performing specific actions activates participants' folk psychological thinking (or "theory of mind"), and this makes people much less likely to think that determinism or mechanism involves bypassing of an agent's beliefs, desires, and decisions. Although the abstract mechanistic cases may prime most people to take what Daniel Dennett (1987) calls the "mechanistic stance" or what Peter Strawson (1962) calls the "objective attitude" towards the agents, both the psychological cases and the concrete cases may prime most people to take the "intentional stance" or the "personal attitude" towards the agents. Taking this stance involves thinking in terms of the agent's desires, beliefs, and reasons, whereas the mechanistic stance leads people to think more in terms of mechanisms that bypass these psychological states.

These results are plausibly interpreted to support the proposed error theory for apparent incompatibilist intuitions: It is not that people found the determinism involved in the scenarios to be incompatible with free will—most did not—it is rather that people found the mechanism in the scenarios to be incompatible with free will because it primes bypassing intuitions (though much less so when bypassing intuitions are dampened by consideration of specific agents and decisions). The deterministic elements in all scenarios were identical. So, we must look elsewhere for an explanation for people's significantly different responses to the different scenarios. Whether the scenario primes people to interpret it to involve bypassing is a plausible place to look. However, we did not ask participants in this study whether they were interpreting the scenarios to involve bypassing, and hence we could not directly compare their judgments about bypassing with their judgments about free will and responsibility. The following studies took this next step.

Measuring Bypassing

Dylan Murray and I used four different scenarios describing determinism, two of which used Nahmias, Morris, Nadelhoffer, and Turner's (NMNT) "rollback" description of determinism (as presented above) and two of which used the description of determinism from Nichols and Knobe [(N&K) 2007] (see below and see chapter 28 for full description). For each description of determinism, we had an abstract case and a concrete case with a specific agent performing a specific negative action (in NMNT, Jill stealing a necklace; in N&K, Bill killing his wife and children). Each participant read one of these four scenarios (N&K abstract, NMNT abstract, N&K concrete, or NMNT concrete) and then answered a series of randomly ordered questions, some of which asked about whether a person in the scenario can have free will, be morally responsible, and deserve blame, and some of which asked about bypassing—for instance, three of the questions were about whether what a person wants (believes, decides) has no effect on what they end up doing.[11] Our main prediction was that a participant's judgments about free will and responsibility would inversely correlate with his or her judgments about these bypassing questions. That is, the more one interpreted a description of determinism to involve bypassing, the less one would attribute free will, responsibility, and blame to the agents, and conversely, the less one interpreted a description of determinism to involve bypassing, the more one would attribute free will, responsibility, and blame to the agents. We were surprised by how strongly the results supported this prediction.

We also predicted that people would be more likely to interpret scenarios to involve bypassing in the abstract cases than the concrete cases. As suggested above, the hypothesis is that specific agents, decisions, and actions are more likely to engage thinking in terms of psychological causes (e.g., desires, beliefs, and reasons), such that people given concrete cases will be less likely to interpret these psychological processes as being bypassed. And given our main prediction, we therefore predicted that participants' judgments about free will, responsibility, and blame would be higher in the concrete scenarios than the abstract scenarios.

Finally, we predicted that the way determinism was described would influence people's judgments about bypassing and their judgments about free will and responsibility. Specifically, we thought that Nichols and Knobe's description of determinism, in the abstract case, would trigger particularly high bypassing judgments, and that this might explain the finding that most people (86 percent) in their original study responded that, in the abstract case, determinism precludes being fully morally responsible. Nichols and Knobe interpreted their results to suggest that people do in fact have incompatibilist intuitions when they are thinking about the issues abstractly or theoretically, but that people tend to say that determinism does not conflict with responsibility when they read the emotionally charged concrete cases—indeed, 72 percent said that Bill was fully morally responsible for killing his wife and children. Nichols and Knobe (2007, 672) suggest that these seemingly compatibilist responses are "performance errors brought about by affective reactions. In the abstract condition, people's underlying theory is revealed for what it is—incompatibilist." I agree that high affect can bias people's responsibility judgments, especially when heinous crimes provoke extreme emotions. However, I also believe that people are more likely to interpret descriptions of determinism to entail bypassing in the abstract cases than the concrete cases, and that this helps to explain why they are more likely to say that agents lack free will and responsibility in the abstract cases. I think this explanation is particularly apt for Nichols and Knobe's results. On the one hand, their concrete case involves a very negative action that is likely to engage not only the intentional stance but also intense affective reactions (e.g., moral outrage), which may indeed bias people to attribute responsibility where they otherwise would, and perhaps should, not. On the other hand, I believe that their abstract case has problematic features likely to elicit bypassing judgments.[12]

Nichols and Knobe's scenario describes two types of universe, one (Universe A) meant to be deterministic, in which "everything that happens is completely caused by whatever happened before it," including decisions, and one (Universe B) in which almost everything is completely caused but in which "the one exception is human decision making." The scenario emphasizes (and ends by saying) that in Universe A "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" (emphases in original). It is not obvious why the wording of the N&K scenario should induce high bypassing judgments, but it does. Two-thirds of our participants who read their abstract scenario agreed with the bypassing statements.[13] I believe that the best explanation for this effect is that the "has to happen" language, as worded, is easily interpreted to mean that everything has to happen *no matter what*. That is, this way of describing determinism may be interpreted as a form of fatalism, such that what one is caused to do must happen *regardless* of what one wants, believes, or tries to do, which suggests bypassing. Indeed, in addition to the high levels of agreement on the four questions used in our bypassing composite score, most participants agreed with other statements that suggest they understood the N&K abstract scenario to entail bypassing: 61 percent agreed with the statement, "In Universe A, everything that happens *has to* happen, even if what happened in the past had been different" (which determinism, properly understood, does not entail); 70 percent agreed with the statement, "In Universe A, a person's trying to get what they want makes no difference to whether they end up getting it"; and 91 percent agreed with the statement, "In Universe A, everything that happens to a person will happen no matter what." In the other three scenarios, far fewer participants expressed agreement with similar "fatalistic" questions.

Each of the three predictions described above were strongly supported by statistical analyses of the results. The N&K abstract case elicited significantly higher judgments of bypassing and significantly lower judgments of free will and responsibility than the NMNT abstract case, and both abstract cases elicited higher bypassing judgments and lower judgments of free will and responsibility than the concrete cases. Most important, the relationships between bypassing judgments and judgments of free will and responsibility were very closely related across all four cases. This can be seen in the results presented in Figure 1, but statistical analyses confirm how strong the relationship is. For instance, in each of the four cases, there were highly significant inverse correlations between composite bypassing scores and composite scores based on participants' highly intercorrelated judgments about free will, moral responsibility, and blame. Collapsing across all four surveys, the correlation coefficient between these composite scores was strikingly high: $r(247) = -0.734$, $p < .001$. Furthermore, the effect that the abstract descriptions of determinism had on people's judgments about free will and responsibility was mediated by whether or not they took the description to involve bypassing (as shown by a mediation analysis)—that is, whether or not someone took the scenario to rule out free will and moral responsibility was primarily caused by whether or not they read the scenario to involve bypassing (see Nahmias and Murray, forthcoming).

These results further support the error theory for apparent incompatibilist intuitions. Because determinism does not in fact entail bypassing—e.g., determinism does not mean that agent's desires, beliefs, and decisions have no effect on what they do—these studies suggest that it is when people misunderstand determinism that they are likely to see it as incompatible with free will. Conversely, when people properly understand that determinism does *not* mean that one's mental activity

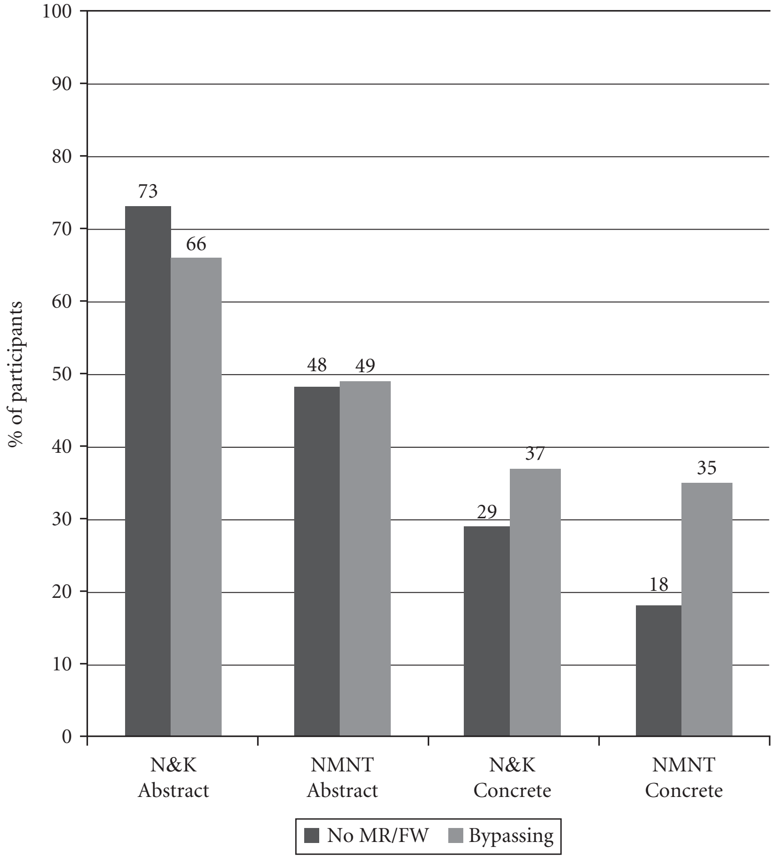

Figure 1. Judgments about MR, FW, and Bypassing
Percentage of 'apparent incompatibilists' (participants with scores indicating disagreement on questions about free will, moral responsibility, and deserving blame) and percentage of 'bypassers' (participants with scores indicating agreement on questions about bypassing: decisions, desires, beliefs have no effect on what happens and agent has no control).

makes no difference to what happens (i.e., that determinism does not entail epiphenomenalism or fatalism), they tend not to take it to rule out free will or responsibility.

But because people can seemingly be so easily confused about what determinism means, depending on how it is described, one might wonder whether we should trust their responses to these surveys as useful information about the intuitions that should matter for the philosophical debates. One might worry that, just as some people misunderstand determinism to entail bypassing, perhaps some people do not really understand the implications that determinism does have. For instance, perhaps the "apparent compatibilists" do not really understand that determinism *does* entail that it is *not* possible, *given* the actual past and laws, for a person to decide otherwise. Finally, one might wonder what happens when people are explicitly told that determinism does not entail bypassing. Murray and I addressed some of these issues in our follow-up studies.

Testing "Competent" Folk

These follow-up studies were designed, in part, to test whether participants comprehend the deterministic element of the scenarios properly and then to see what the "competent" participants (i.e., those who did understand the deterministic element) said about the freedom and responsibility of agents in those scenarios. In all of my studies, we remove from analyses all participants who miss basic comprehension questions, indicating that they did not carefully read and understand the scenarios.[14] These studies take the further step of examining whether participants were responding in ways that suggest that they understand, on the one hand, that determinism does not entail bypassing, and on the other hand, that determinism does entail that a particular past and laws guarantee a particular future.

Our participants first read either Nichols and Knobe's abstract scenario or Nahmias, Morris, Nadelhoffer, and Turner's abstract scenario, but with a section added to emphasize that determinism does not entail bypassing. For instance, after describing the deterministic nature of Universe A relative to Universe B using the original wording, the N&K scenario concludes:

> The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision. This does *not* mean that in Universe A people's mental states (their beliefs, desires, and decisions) have no effect on what they end up doing, and it does *not* mean that people are not part of the causal chains that lead to their actions. Rather, people's mental states *are* part of the causal chains that lead to their actions, though their mental states are always completely caused by earlier things in the causal chain that happened before them—*given* that the past happened the way it did, each decision *has to happen* the way 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 given what happened in the past.

And the NMNT scenario concludes:

> This does *not* mean that in Universe C people's mental states (their beliefs, desires, and decisions) have no effect on what they end up doing, and it does *not* mean that people are not part of the causal chains that lead to their actions. Rather, people's mental states *are* part of the causal chains that lead to their actions, though their mental states are always completely caused by earlier things in the causal chain that happened before them—if a person decides to do something in Universe C, then *every* time the universe is re-created with the *same* initial conditions and the *same* laws of nature, that person decides to do the same thing at that time and then does it.

As predicted, these additions led significantly fewer participants to agree to the bypassing questions than in the previous study. However, some still misinterpreted determinism to involve bypassing, and these participants were significantly more likely to offer apparent incompatibilist intuitions than those who did not. For instance, in N&K abstract, 49 percent still had a composite score indicating agreement to the bypassing questions, and of these, 70 percent were "apparent incompatibilists," with composite scores indicating disagreement with statements about agents' having free will, being morally responsible, and deserving blame. If we then treat agreement with bypassing questions as a misunderstanding of the scenario and remove these participants from analysis, we find that only 32 percent of the remaining subjects gave incompatibilist judgments (compare this, for instance, to the 86 percent of participants who gave such judgments in Nichols and Knobe's original study). In NMNT abstract, fewer participants missed the bypassing questions (29 percent), though of those, most also judged that agents lack free will and responsibility (85 percent), again indicating a strong relationship between bypassing judgments and apparent incompatibilist judgments. When we remove those who missed the bypassing questions, only 18 percent gave incompatibilist judgments.

But what happens when we remove participants who do not seem to grasp that determinism "fixes" the relationship between a particular past and particular future events? To get at this issue, in addition to removing those who missed comprehension questions about the deterministic elements of the scenario (see note 14), we asked participants a "modal question." Those reading N&K abstract were asked whether they agreed with this statement: "In Universe A, *given* that past events happen the way they do, it *has to happen* that later events happen the way they do."[15] Most (90 percent) answered correctly (i.e., agreed). Of the few who answered incorrectly, most offered what we can call "apparent compatibilist intuitions." Nonetheless, if we remove those who missed this modal question (as well as those who missed the bypassing questions), we find that the majority of remaining "competent" participants in N&K abstract still offered compatibilist judgments (62 percent). Likewise, in NMNT abstract, participants were asked whether they agree that "In Universe C, if the universe is re-created with the exact *same* initial conditions and laws of nature, then it *has to happen* that later events happen the way they do." Again, most (81 percent) answered correctly (i.e., agreed), and the majority of those missing it offered apparent compatibilist judgments. If we remove those who missed this modal question and those who

Table 3 Summary of Results of "Competent" Participants (i.e., those who answer bypassing and modal questions correctly)

Percentage of participants with composite score indicating…	**N&K Abstract**	**NMNT Abstract**	**N&K Concrete**	**NMNT Concrete**
…agreement on questions about **free will**, **moral responsibility**, and deserving **blame**	62%	78%	89%	89%

missed the bypassing questions, again we find that the majority (78 percent) of the remaining "competent" participants offered compatibilist judgments (see Table 3).

Participants then read the scenarios with an added paragraph describing a specific agent performing a bad action, which concluded: "At a particular time one day (time T), Bill decides to enter the store and steal the necklace and he then does so." As predicted, priming participants to think about a specific person and action led fewer of them to miss the bypassing questions and more of them to attribute free will, responsibility, and blame to the agent, and of those who did not miss the bypassing questions, very few (10 percent) offered incompatibilist responses. A few more participants missed the modal questions[16] in these concrete cases than in the abstract ones, but if we remove them (along with those who missed the bypassing questions), we find that an even higher proportion of these "competent" participants offered compatibilist responses than in the abstract cases: a full 89 percent in both N&K and NMNT (see Table 3). That is, when considering particular agents in a deterministic universe, *9 out of 10* of the "competent" participants had composite scores that indicate they agree that the agent acts of his own free will, is fully morally responsible for his action, and deserves to be blamed for it.

Given the body of research I have presented and barring countervailing evidence, I think we can conclude several things about people's intuitions about free will and moral responsibility. First, depending on how determinism is described, people often express *apparent* incompatibilist intuitions, interpreting determinism as a threat to free will because they misinterpret it to involve bypassing. These intuitions help to explain the tendency to treat compatibilism as counterintuitive. There are also some apparent compatibilists, people who do not see determinism as a threat to free will perhaps only because they do not understand that determinism does impose some "modal constraints" on agents. However, among the remaining "competent" participants, the genuine incompatibilists (who see determinism as a threat to free will even though they recognize that it does not involve bypassing) represent a very small minority of the folk. There are many more genuine compatibilists who do not find determinism threatening to free will even though they understand the modal constraints it imposes on agents. Finally, the evidence suggests that people are more likely to express compatibilist intuitions—and are less likely to make bypassing errors—when they are considering concrete cases rather than abstract cases, presumably because they are primed to take the intentional stance and the participant stance towards the agent in question. My own view is that the intuitions that are most relevant to debates about free will and moral responsibility are not those that are generated when people are considering abstract cases in a more theoretical or detached way, but instead those that are generated when people are considering specific agents and actions, thus engaging "theory of mind" capacities (though not when considering agents or actions so abhorrent that they engage emotional biases).[17]

Why Compatibilism does not Secure Free Will

Hence, the overall picture suggested by the data is that incompatibilism is not more intuitive than compatibilism, that compatibilists are not propping up a "wretched subterfuge" and do not have the burden of motivating a counterintuitive position, and that one reason incompatibilism might *appear* intuitive is that determinism is, at least among the folk, often interpreted to entail bypassing. I take these results of experimental philosophy studies to have significant implications for the free-will debates. They do not, of course, demonstrate that compatibilism is true. But if only a minority of people have genuine, rather than merely apparent, incompatibilist intuitions—and especially if those intuitions are not deeply entrenched—then compatibilism would be at most a slightly revisionary theory rather than a significant conceptual revolution. And if people's ordinary intuitions do not provide support for incompatibilism, it is hard to see what should motivate revision towards an incompatibilist theory whose metaphysical demands leave human free will in a rather tenuous position. On a libertarian theory, it could turn out that physicists' discovering that determinism is true would suddenly reveal that people do not have, and never have had, free will or moral responsibility (assuming it requires free will). Such a theory also suggests that we may be in no position to know whether or not humans have free will (and are morally responsible), assuming we are not in a position to know whether determinism is true (or to know whether humans have the proper agent-causal powers).

This worry that incompatibilism leaves human free will in such a tenuous position may be what has motivated some to defend compatibilism, because as Kane (2005a, 13) suggests, compatibilism seems to allow that we "need not worry that future science will somehow undermine our ordinary conception that we are free and responsible agents." But compatibilism is a thesis about possibilities (whether free will and determinism are compossible), not about actualities (whether humans actually have free will). One might think that compatibilism secures human free will if one thinks that determinism is the *only* possible threat to human free will. Historically, most compatibilists seem to have thought that way. But clearly nonphilosophers do not think that way. As we have seen, they seem more concerned about bypassing threats than determinism. And if bypassing were true—for instance, if our rational mental processes did not play the proper role in our decisions and actions—that would threaten free will on just about every theory, compatibilist or incompatibilist. Bernard Berofsky (in Taylor 2005, 82) makes this sort of point when he writes:

> All parties to disputes about freedom and autonomy must agree that a necessary condition of the very possibility of freedom and autonomy is that we act as we do for the reasons we cite.... Both [compatibilist and incompatibilist] ought to be driven by the thought that free and autonomous agents are responsive to reasons

> in a sense that precludes an account of behavior in terms of neurophysiological processes that displace the one in terms of reasons.

There may be several ways that, in theory, bypassing could turn out to be true (e.g., eliminativism or epiphenomenalism in philosophy of mind). But some scientists are already claiming that bypassing *is* true. For instance, some think that neuroscientific evidence shows that our brains make decisions before we are consciously aware of having made them, such that our conscious mental processes play no causal role in action (e.g., Libet 1999; Soon, Brass, Heinze, and Haynes 2008). Some think that psychological evidence demonstrates that conscious will is an illusion (e.g., Wegner 2002) or that our rational mental processes do not influence our judgments and actions but only come up with rationalizations for them after the fact (e.g., Bargh 2008). Finally, some seem to assume that a scientific explanation of human behavior entails that (conscious) mental processes are epiphenomenal such that free will is an illusion (e.g., Montague 2008). (See the essays of Mele and Walter in this volume for further discussion of many of these views.)

My own view is that the relevant scientific evidence does not in fact support these bypassing challenges and that such claims are often based on conceptual confusions, though I do think the evidence about rationalization suggests that we possess *less* free will than we tend to think (see Nahmias 2010). My point here, however, is to highlight that these bypassing challenges have been put on the table, and they are distinct from the supposed threat of determinism. Crucially, ordinary people are much more likely to recognize and worry about such bypassing challenges than determinism. Because the scientific claims are increasingly being publicized, and because the public is "intuitively poised" to take them to heart, philosophers on all sides of the free-will debate should make sure that their myopic focus on determinism does not blind them to other potential threats to free will.[18]

NOTES

1. This chapter draws significantly on work I have published with former students and co-authors Thomas Nadelhoffer, Stephen Morris, Jason Turner, Trevor Kvaran, Justin Coates, and Dylan Murray. I consider all of them collaborators on this chapter, though they do not all agree with my conclusions. I thank each of them for all the assistance and stimulation they have provided me in my experimental philosophy research. I would also like to thank the thousands of participants who have taken our surveys and helped me learn more about what people think about free will and moral responsibility. For helpful comments on this chapter, I thank Jason Turner, Joshua Knobe, and especially Dylan Murray. This chapter was completed in part with support from a grant from the University of Chicago Arete Initiative and the John Templeton Foundation.

2. Some other incompatibilists who claim their view is intuitive include Ekstrom (2002, 310), Pereboom (2001, xvi), Pink (2004a, 12), and Cover and Hawthorne (1996, 51),

not to mention, as Kane points out, William James and Kant. Compatibilists who suggest their position is intuitive include Dennett (1984), Wolf (1990, 89), Lycan (2003), and Nowell-Smith (1949, 49).

3. Consider, for instance, the way Richard Taylor (1974, 36) describes the consequences of determinism in an oft-anthologized piece: "What am I but a helpless product of nature, destined by her to do whatever I do and to become whatever I become?" He then claims that there is no difference between "an ingenious physiologist [who] can induce in me any volition he pleases" and "perfectly impersonal forces" such as deterministic laws. He continues, "Whether a desire which causes my body to behave in a certain way is inflicted upon me by another person, for instance, or derived from hereditary factors, or indeed from anything at all, matters not in the least" (46).

4. In Nahmias, Morris, Nadelhoffer, and Turner (2004), we also discuss the importance of empirically studying people's phenomenology of free will, such as their experiences of deliberating, making choices, or exercising self-control. Philosophers with different theories tend to disagree about the relevant "folk phenomenology" as much as they disagree about the relevant folk intuitions and folk theories.

5. We also had a neutral scenario in which Jeremy goes jogging, in response to which most people said that he had free will. All results presented in Table 1 are significantly different from chance, as determined by χ^2 goodness-of-fit tests. Participants in these scenarios were undergraduates at Florida State University. For complete scenarios, questions, methods, and results, see Nahmias, Morris, Nadelhoffer, and Turner (2005, 2006).

6. The wording of the scenario as presented here includes minor revisions of the one used by Nahmias, Morris, Nadelhoffer, and Turner (2006) in order to reflect the exact wording used in the study presented below in the section, "Measuring Bypassing."

7. See Nahmias, Morris, Nadelhoffer, and Turner (2006, section 4), for further discussion of these issues.

8. Underlying these intuitive responses may be a tendency to assume (presumably implicitly) that a complete causal explanation of phenomena Z in terms of X can leave no room for any other causal contribution from Y. In the "horizontal" direction, this "single explanation assumption" (SEA) takes the form: If X at time t1 is causally sufficient for Z at time t3, then Y at time t2 plays no causal role in bringing about Z. Of course, this is poor reasoning when X brings about Z *by causing* Y to cause Z (e.g., lighting the fuse makes the bomb explode by causing the fuse's burning to detonate the bomb). In the "vertical" direction, SEA may take a form that looks something like Jaegwon Kim's (1998) "causal exclusion argument": if Y supervenes on X (such that X is sufficient for Y) and if X is causally sufficient for Z, then Y could only causally contribute to Z by overdetermination.

9. Bypassing might help to explain the intuitive appeal of conditional analyses of the ability to do otherwise. Such analyses say that if the agent had thought, wanted, or decided differently, then she would have acted otherwise, which entails that bypassing is false if we take bypassing to mean that the agent would have done what she did *regardless* of what she had thought, wanted, or decided.

10. For more information on methods and results, see Nahmias, Coates, and Kvaran (2007). Differences described in the text were statistically significant as demonstrated using ANOVAs and t-tests. Participants in this study, as well as the ones described below, were undergraduates at Georgia State University in Atlanta, Georgia.

11. For details about these studies, see Nahmias and Murray (forthcoming). Analyses were performed on 249 participants at Georgia State University, distributed roughly equally across the four cases.

12. Notice that Nichols and Knobe's "affective biasing" interpretation has difficulty explaining the results of Nahmias, Morris, Nadelhoffer, and Turner's results in the cases that involved positive actions (e.g., an agent's saving a child, returning a wallet, or giving to charity) and neutral actions (e.g., an agent's going jogging). For more detailed critiques of Nichols and Knobe's studies, see Nahmias (2006), Turner and Nahmias (2006), and Nahmias and Murray (forthcoming).

13. More precisely, 66 percent had a "composite bypassing score" above the midpoint on a six-point scale indicating agreement, where the composite was created by combining responses to four questions that were very highly intercorrelated: whether or not (1) what a person wants has no effect on what they end up being caused to do; (2) what a person believes has no effect…; (3) a person's decisions have no effect…; and (4) a person has no control over what they do. Though (4) correlates very closely with (1–3) and removing it from the statistical analyses has little effect, some may dispute that the "no control" question is an appropriate measure of bypassing judgments.

14. For instance, in the studies described in the previous section, we removed from analysis participants who missed either of two true/false questions, such as: "According to the scenario, in Universe C the same initial conditions and the same laws of nature cause the exact same events for the entire history of the universe." We also removed participants who completed the survey significantly more quickly (and presumably more carelessly) than the average participant. For these follow-up studies, the number of remaining participants, whose responses are described below, was 141.

15. Note that this question and the one described later in this paragraph are still potentially problematic in that the scope of the modal operator is ambiguous. With this worry in mind, we worded the questions in the clearest way we could, while also being consistent with the wording used in the scenarios (see Turner and Nahmias [2006] for discussion of these modal scope issues).

16. For instance, in N&K concrete, participants are asked whether they agree or disagree with this statement: "*Given* everything that happened before time T, Bill *has to* decide to steal the necklace at time T."

17. In general, philosophers seem to trust and appeal to intuitions generated by concrete cases more than intuitions generated by abstract cases or questions. For instance, epistemologists tend to take intuitions offered in response to Gettier cases to be more informative and reliable than intuitions offered in response to abstract questions about whether an agent knows that *p* if she believes that *p* on the basis of good evidence and *p* is true.

18. Knobe and Nichols' insightful discussion in the previous chapter (28) similarly suggests that people's intuitions about the self and free will are not sensitive to determinism *per se* but rather to bypassing of certain features of the self. I agree with their general approach, though I think it is more accurate to treat people's conception of the 'executive self' as a subset of the 'psychological self' rather than as a distinct entity or level of analysis, one that, for instance, commits them to substance dualism or agent causation. Depending on people's perspective (or 'level of zoom'), they may treat some desires or emotions as external to the self, but this does not entail that they treat a person's conscious deliberations, reflective beliefs, and endorsed desires as part of a *non-psychological* self (and I don't think Knobe and Nichols' experimental results support this model for a folk theory of a non-psychological, or supra-psychological, executive self).

Combining our views, it may be that what people count as bypassing depends on their 'level of zoom', but typically it is bypassing of specific features of an agent's psychological self, broadly construed to include conscious and self-reflective mental processes, that will be perceived as a threat to free will. It is precisely because modern neuroscience is often presented as demonstrating that our brains cause behavior in a way that bypasses these conscious mental processes that it is taken to threaten free will. I predict that most people will not interpret neuroscience to threaten free will if neuroscience (plus a good philosophical and scientific theory of consciousness) turns out to help *explain* how conscious mental processes work, rather than explaining them *away*.