Moral Psychology

Volume 4: Free Will and Moral Responsibility

edited by Walter Sinnott-Armstrong

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

1 Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences

Eddy Nahmias

Questions about free will and responsibility have long been considered the purview of philosophers. If philosophers paid attention to any science, it was physics since physics might tell us about whether or not the traditional threat of determinism is true. This is changing, though too slowly. Philosophers considering human autonomy and responsibility need to pay more attention to the relevance of the sciences that study humans, in part because neuroscientists and psychologists are increasingly discussing free will, usually to argue that their research shows that it is an illusion. For instance,

> Neuroscientist Patrick Haggard says: "We certainly don't have free will. Not in the sense we think." (in Chivers, 2011)
>
> Psychologist John Bargh (2008) writes, "The phenomenological feeling of free will is very real ... but this strong feeling is an illusion, just as much as we experience the sun moving through the sky, when in fact it is we who are doing the moving." (pp. 148–149)
>
> Psychologist Daniel Wegner (2002) concludes *The Illusion of Conscious Will*: "It seems we are agents. It seems we cause what we do.... It is sobering and ultimately accurate to call all this an illusion." (pp. 341–342)
>
> Cognitive scientists Joshua Greene and Jonathan Cohen (2004) argue, "The net effect of this influx of scientific information will be a rejection of free will as it is ordinarily conceived with important ramifications for the law." (p. 1776)
>
> Neuroscientist Sam Harris (2012) writes, "Free will *is* an illusion. Our wills are simply not of our own making.... your brain has already determined what you will do." (pp. 5, 9)

Some of the other scientists who have suggested that research in neuroscience and psychology threatens the existence of human free will include Francis Crick (1994), Benjamin Libet (1999), Mark Hallet (2007), Sue Pockett (2007), Read Montague (2008), Anthony Cashmore (2010), and Stephen Hawking (2010).

These claims get a lot of play in the media, in part because they are eye-catching. A headline in London's *Sunday Times* (10/21/09) reads, "Sexy Science: Is Free Will Just an Illusion?" A *ScienceNews* (12/6/08) article reports, "'Free will' is not the defining feature of humanness, modern neuroscience implies, but is rather an illusion that endures only because biochemical complexity conceals the mechanisms of decision making." And Jerry Coyne writes in *USA Today* (1/1/12):

> The debate about free will, long the purview of philosophers alone, has been given new life by scientists, especially neuroscientists studying how the brain works. And what they're finding supports the idea that free will is a complete illusion. The issue of whether we have of free will is not an arcane academic debate about philosophy, but a critical question whose answer affects us in many ways: how we assign moral responsibility, how we punish criminals, how we feel about our religion, and, most important, how we see ourselves—as autonomous or automatons.[1]

I will argue that Coyne and the other scientists challenging free will are mistaken about what the science actually shows. However, I agree with Coyne that these debates matter. Our beliefs about free will influence our self-conception and our moral and legal practices. Recent research has also shown that when people are told that science shows free will is an illusion, it temporarily influences their behavior, for instance, leading them to cheat more, help less, act meaner, exert less self-control, think less about alternatives, and make less punitive judgments (Vohs & Schooler, 2008; Evans, in press; Baumeister et al., 2009; Baumeister, this volume).

Because of the practical implications of people's beliefs about free will, it is crucial that we properly understand what scientific discoveries actually reveal about free will. And because scientific claims about free will are being widely publicized, it is increasingly important to ensure that these claims match up with what people actually believe about free will.

Imagine an Imaging Study

One way to tell what people actually believe about free will is to ask them about possible cases. Imagine people read this story in a reputable science news publication:

> Your decision to read this story was carried out entirely by your brain. In a study published in *Nature Neuroscience*, researchers using brain scanners could see exactly which brain processes occurred as people made decisions, and they found earlier brain activity that correlated with the decisions people would make.
>
> "We have discovered that our decisions are caused entirely by the complex processes happening in the brain," says Peter Bernstein at the Center for Neuroscience at Princeton University. In his study, students were shown descriptions of three psychology courses, considered reasons for and against each of them for up to one minute, and then pressed one of four buttons indicating their decision to sign up for one of the courses (or none of them).
>
> All of this occurred while they were lying in a new type of functional magnetic resonance imaging (fMRI) scanner, which is able to measure where and when brain activity occurs, as well as the correlations between specific brain activity and other brain activity. The researchers were able to measure how earlier brain processes, such as the ones involved in the participant's conscious memories and desires, provided information about the later ones, including the decision itself.
>
> For instance, Bernstein explained, "One participant was a young woman who had already taken two of the courses and was not interested in the third course. We could see the brain processes that corresponded to her memory of the previous courses as they caused the processes that corresponded to her conscious experience of disliking those courses. And as she read the third description, we could see the processes that corresponded to her negative reaction to that course. We were able to use the information about this earlier brain activity to predict her decision to push the fourth button for 'none of the above' with 70% accuracy."

This brain imaging study is imaginary. Neuroscientists are not yet able to map the neural activity involved in such complicated and extended decision-making tasks, and 70% accuracy in predicting choices among four options has not yet been achieved by any real study. Existing studies deal with much simpler decisions, such as Libet's (1985) infamous study on voluntary wrist flexes or recent extensions of his paradigm, in which participants decide whether to push a left or right button. For instance, John-Dylan Haynes and colleagues had people repeatedly make this left–right decision in an fMRI scanner and found patterns of neural activity 7 to 10 seconds before the button press that "predicted" the participant's decision; more specifically, an impressive new form of data analysis found correlations at 10% above chance between patterns of brain activity in frontopolar cortex and which button was pressed. The authors suggest that this discovery supports the Libet-inspired conclusion that the "subjective experience of freedom is no more than an illusion and that our actions are initiated by unconscious mental processes long before we become aware of our intention to act" (Soon et al., 2008, p. 543; Haynes, this volume). Haynes laments, "I'll be very honest, I find it very difficult to deal with this.... How can I call a will 'mine' if I don't even know when it occurred and what it has decided to do?" (in Smith, 2011, p. 24).

If this button-pressing study is "difficult to deal with," then the imaginary fMRI study should be more distressing. In fact, however, when people read about it, their interpretation depends crucially on how the scientists

present its implications. In a pilot study I ran with 152 participants, one group read the study information above with a headline that read, "Neuroscientists Discover that Free Will is an Illusion," and with quotations from the (fictional) neuroscientist Bernstein such as, "Our brain causes our decisions and then we consciously experience the outcome, much like a spectator observing a play." A second group read the exact same information about the study but with a headline that read, "Neuroscientists Discover How Free Will Works" and with quotations from Bernstein such as, "By understanding these complex processes in our brains, we are understanding how conscious deliberation and self-control work."

In the first group, 63% responded that the study provides either some, strong, or convincing evidence that people do *not* have free will (and the majority responded that it provides evidence that what people do is *not* really up to them and that people do *not* control what they do). However, in the group that read the scientist's more optimistic assessment of the exact same study, only 16% responded that it provides evidence against free will while 68% responded that the study provided either some, strong, or convincing evidence that people *do* have free will (and over 60% responded in that way to the "up to" and "control" questions).[2]

One way to read these results is that people are pushovers and just go along with whatever scientists say. Another possibility is that studies like the imaginary one described here do not actually provide clear evidence either for or against free will. Their relevance to free will depends largely on how they are interpreted and presented—and as we'll see below, they can be properly interpreted as evidence that helps to explain how free will works. The way scientists present their interpretations of such studies can clearly impact people's beliefs about free will, and it can also impact behavior (Vohs & Schooler, 2008; Baumeister et al., 2009; Baumeister, this volume). It is not yet clear what drives these behavioral effects related to diminishing people's belief in free will (Nahmias, 2011c) nor how long-lasting they might be. It also remains an open empirical question what practical effects might follow from a more pervasive change in people's beliefs about free will and responsibility. Some skeptics about free will predict negative consequences to our self-conception so severe that we should hide the truth from people (Smilansky, 2000), while other skeptics suggest it should and would have many positive effects, especially by undermining retributive justifications for punishment (e.g., Pereboom, 2001; Greene & Cohen, 2004; Harris 2012).

Either way, changes in beliefs about free will, in people's behavior, and in society's moral and legal practices are most likely to be induced with

the authority of science and by the sorts of claims that are increasingly promulgated by scientists and disseminated in the popular press. For better or worse, it clearly matters whether the scientists are right that their discoveries challenge free will. Are they right?

Do Scientific Discoveries Challenge Free Will and Responsibility?

To interpret potential challenges to free will and responsibility from the modern sciences of the mind, we can use this argument schema:

1. Free will requires that *X* is *not* the case.
2. Science is showing that *X* is the case (for humans).
3. Thus, science is showing that humans lack free will.

Furthermore, assuming (as I and many others do) that free will involves the capacities that allow agents to be morally responsible for their choices and actions—for instance, to *deserve* credit and blame and to be appropriate targets of reactive attitudes such as indignation and gratitude (Strawson, 1962)—then this argument would further conclude that

4. Science is showing that humans are not morally responsible agents.[3]

This schema illustrates how much depends on the "*X* factor" in premises 1 and 2. To work, the argument needs some replacement for "*X*" that makes both premises true together. If scientists target an *X* factor that does not match what is properly required for free will, then not only will their conclusion be unjustified but they also risk influencing people to believe they lack free will when in fact they only lack what scientists mistakenly believe free will requires. As it turns out, the scientists are often ambiguous about what is supposed to fill in this argument schema—so, let's consider various options.

Determinism

Perhaps the *X* factor is supposed to be "determinism." After all, a prominent philosophical theory of free will, *incompatibilism*, says:

1D. Free will requires that *determinism* is not the case.[4]
If 1D is true, then human free will and responsibility would be undermined by this premise:
2D. Science is showing that determinism is the case (for humans).

Indeed, Bargh and Ferguson (2000) argue that they will "present the case for the determinism of higher mental processes" (p. 926) in order to reach

the conclusion that free will is an illusion, and many other scientists use "determinism" to describe the challenge they think neuroscience and psychology pose to free will (e.g., Libet, 1999; Tancredi, 2007).

An initial problem with this way of posing the challenge is that the philosophical arguments advancing premise 1D define determinism differently than cognitive scientists seem to understand it. In incompatibilist arguments, determinism is defined as the thesis that a complete description of a system (e.g., the universe) at one time and of all the laws that govern that system logically entails a complete description of that system at any future time (e.g., van Inwagen, 1983). However, research in the cognitive sciences is simply *not* in a position to establish determinism, so defined. Determinism requires a closed system, but scientists who study human brains and behavior do not study closed systems. Furthermore, none of the specific discoveries touted as challenging free will, such as Libet's, Haynes's, or Bargh's, help to establish the truth of determinism since they do *not* show that, given prior events (e.g., specific neural processes or psychological manipulations), certain decisions or behaviors *necessarily* occur (see Roskies, this volume).[5]

Of course, it's entirely unclear how indeterminism at *any* level could help secure free will. While some philosophers have looked to quantum indeterminism in the brain to allow for free will (Kane, 1996), most have concluded that such indeterminism alone could not provide us with any relevant type of control or responsibility that we could not have without such indeterminism. This should make us wonder why philosophers have focused so much attention on determinism and whether they have neglected scientific discoveries more relevant to human free will—for instance, discoveries about how human minds actually work, rather than discoveries about the fundamental laws of physics.

In fact, most contemporary philosophers are *compatibilists* who do not believe determinism is relevant to free will. They reject premise 1D, arguing that determinism does not conflict with an agent's possessing the cognitive capacities required for free will or responsibility, such as the capacity to govern one's decisions and behavior in light of one's reasons, nor does it conflict with our abilities or opportunities to exercise those capacities. There are numerous well-developed theories of free will that make the issue of determinism entirely irrelevant to whether agents can have free will. Scientists cannot simply assume that determinism, properly understood, rules out free will.[6]

Nor can scientists assume that ordinary people understand determinism to rule out free will. When Haggard claims we do not have free will "in the sense we think," or Greene and Cohen (2004) conclude, "Free will, as we ordinarily understand it, is an illusion" (p. 1783), they highlight a common maneuver: to define free will in terms of armchair assumptions about how most people understand it. Scientists should be eager to consider *empirical* studies of the way people actually think about free will, a task recently taken up by experimental philosophy.

This research in experimental philosophy has discovered some complicated patterns in people's judgments about free will (Nichols & Knobe, 2007; Nichols, 2011). However, my work suggests that most ordinary people do *not* take determinism, properly construed, to threaten free will or moral responsibility. Rather, most people take deterministic scenarios to rule out free will only if they *mis*understand determinism to mean that agents' mental states are "bypassed" such that they do not contribute to action (Nahmias et al., 2006, 2007; Nahmias & Murray, 2010; Murray & Nahmias, 2012). That is, when people read scenarios that describe deterministic universes, most respond that agents in those universes can have free will, be morally responsible, and deserve praise or blame for their actions. Those who reject these possibilities typically do so because they take determinism to mean that the agent's beliefs, desires, and decisions have no effect on what they do. However, determinism does *not* have those implications—mental states can be causally efficacious even if they are deterministically caused. Determinism means that different causes have different effects; hence, if mental states are part of the causal order, different mental states will cause different behavior. For instance, your deciding to do *X* was caused by your caring a lot about *X*; if you had cared less about *X*, that would have caused you to make a different decision. Intuitions that *seem* to support incompatibilism instead indicate that people find free will to be threatened by *bypassing*, the idea that our mental states do not play the proper causal role in our decisions and actions. We will see below that the idea that our mental states are bypassed by neural events seems to be what drives the scientific challenges to free will.[7]

The cognitive sciences can and should contribute to this research on people's understanding of free will and what they take to threaten it. However, research in neuroscience and psychology is not in a position to settle the long-standing debate about the compatibility of free will and determinism, and scientists cannot simply assume that the incompatibilist premise 1D is correct or commonsensical. Nor should they assume that science is establishing that determinism is true (2D)—indeed, the dominant theory of quantum physics suggests it is not.

As it turns out, when neuroscientists and psychologists discuss free will, they do not really seem concerned about the truth or falsity of determinism, in the sense used in traditional debates about premise 1D. Rather, they are using the term more loosely. This means it can be hard to pin down exactly what they have in mind.

Naturalism

One thing some scientists seem to mean by "determinism" is something better described as *naturalism*, the view that everything that exists, including human minds, is part of the natural world and behaves in accordance with natural laws.[8] Naturalism clearly does not entail determinism since quantum indeterminism is consistent with naturalism, and determinism does not obviously entail naturalism since nonnatural souls or minds could still obey deterministic lawful interactions. Haggard clarifies his conclusion by saying this: "We don't have free will, in the spiritual sense. What you're seeing is the last output stage of a machine.... But there's no ghost in the machine" (in Chivers, 2011). Neuroscientist Read Montague (2008) is more explicit:

> Free will is the idea that we make choices and have thoughts independent of anything remotely resembling a physical process. Free will is the close cousin to the idea of the soul—the concept that "you," your thoughts and feelings, derive from an entity that is separate and distinct from the physical mechanisms that make up your body.... Consequently, the idea of free will is not even in principle within reach of scientific description. (p. 584)

And Greene and Cohen (2004) assert that people's conception of free will is "implicitly *dualist* and *libertarian* ... the mind and brain are separate, interacting, entities" (p. 1779). If scientists stipulate this definition of free will, then determinism versus indeterminism is not the issue. Rather, they seem to have in mind an argument like this:

1N. Free will requires that *naturalism* is *not* the case.
2N. Science is showing that *naturalism* is the case.
3N. Thus, science is showing that humans lack free will (and moral responsibility).

It is unclear that any scientific discoveries could conclusively *demonstrate* naturalism (premise 2N) rather than assuming it as a methodological principle. However, science has certainly offered increasing inductive evidence for naturalism, including naturalism about human decision making and behavior, by providing increasingly complete explanations for observable

events in the universe, including human behavior, in terms of natural processes and laws. If nonphysical minds do exist, they seem to have less and less to do. I think that we have good arguments and evidence for naturalism and that we should assume naturalism is true and see how far we can get with it in trying to understand what free will is and how it works. If so, then this argument turns entirely on premise 1N.

The definition of free will used in premise 1N needs to be motivated, in part because it is more metaphysically bloated than naturalistic alternatives. Again, scientists seem to motivate it by assuming that it is demanded by most people's definition of or intuitions about free will, or they think it is accepted philosophical orthodoxy. In fact, among philosophers, very few define free will in such a way that it requires mind–body dualism; instead, among contemporary philosophers, all compatibilists and most incompatibilists develop theories of free will meant to be consistent with naturalism.[9] There are indeed nonnaturalistic conceptions of free will floating around in religious discussions and among some ordinary folk, especially those with specific religious views. However, neither philosophers nor most folk think that free will requires immaterial minds or souls.

My research on nonphilosophers' understanding of free will suggests that a minority of people think that free will depends on a nonphysical mind or soul or that free will would be undermined by naturalism. For instance, in one survey using a representative sample of the U.S. population, almost 400 participants were asked whether they agreed or disagreed with this statement: "If it turned out that people lacked non-physical (or immaterial) souls then they would lack free will." Only 29% agreed, while 41% disagreed, and 30% were neutral (almost identical responses were given to a statement replacing "free will" with "moral responsibility for their behavior"). And only 5% *disagreed* with the statement "People could have free will even if scientists discovered all the laws that govern all human behavior" while 79% agreed and 16% were neutral.[10] Also, recall that very few people who read the fictional study above, without the skeptical interpretation, took it as evidence against free will, even though it emphasizes that neuroscience has discovered that "your decisions are carried out entirely in your brain." They do *not* seem averse to the idea, expressed in the positive version, that "Because the results of this study reveal how decision making works in the brain, the researchers think they have shed light on how free will works." A naturalistic theory of free will would conflict with some people's theory of free will, but most people seem willing to accept that free will is compatible with our minds, in some sense, being our brains, as long as such naturalism is not taken to mean that our

conscious mental states do not matter (see below; see Nahmias & Thompson, in press).

If one stipulates that free will requires a nonphysical soul, then free will would face metaphysical objections, dating back to Descartes, which do not require scientific discoveries to illustrate—namely, explaining how nonphysical minds can causally influence physical bodies. If the challenge to free will is really supposed to derive from naturalism, then no *specific* discoveries from cognitive science do much to advance this challenge. General conclusions about the metaphysics of mind are unlikely to be illuminated with *specific* scientific findings, including the oft-discussed ones about where and when various events happen in the brain.

It seems backward for cognitive scientists to simply assume a nonnaturalistic or dualist theory of free will, since the history of cognitive science can be seen as a series of attempts to demonstrate how we can *put aside* dualistic theories of mind and of cognitive functioning. Descartes argued that humans' cognitive capacities to use language and reason simply could not be explained in terms of natural mechanisms. As cognitive scientists increasingly explain how the mechanisms of the brain can explain language and flexible reasoning, they do not thereby conclude that we *lack* these capacities. Rather, they conclude that dualist theories of such capacities are false.[11] An objective of cognitive science is to find out how the cognitive capacities of the mind/brain work, not to argue that they are illusions *because* they work in nonmagical ways. The sciences of the mind are in a position to *explain* free will, rather than explaining it *away*. Or, as the more optimistic of my fictional neuroscientists suggests, "We have discovered that our decisions are caused entirely by the complex processes happening in the brain. This explains how free will works." Why don't *non*fictional scientists see it this way?

Epiphenomenalism

One motivation for scientists' nonnaturalist view of free will seems to be an assumption that free will requires that conscious mental states play a causal role in behavior, combined with the assumption that consciousness cannot be naturalized. The argument against free will then starts with a more plausible understanding of free will, one that requires that conscious mental processes play an appropriate causal role in our actions, and then it makes the move that science poses a threat by showing that conscious mental processes cannot play a causal role in behavior (i.e., epiphenomenalism). The argument then looks like this:

1E. Free will requires that *epiphenomenalism* is not the case.
2E. Science is showing that epiphenomenalism *is* the case.
3E. Thus, science is showing that humans lack free will (and moral responsibility).

Connecting free will and consciousness (premise 1E) is a good move, though spelling out the connection takes work, as we will see below. But again, it is entirely unclear why cognitive scientists should simply assume that consciousness cannot be naturalized or that it cannot be causally efficacious if it is naturalized.

Suppose we assume that conscious processes can be "naturalized" to the extent that we assume that they have neural correlates; every conscious mental state supervenes on some neural state. If so, then any claim that conscious processes play no causal role in action faces a dilemma. Either one argues that they play no causal role (1) because one assumes that it is their neural correlates that do all the real causal work such that the conscious properties are epiphenomenal, or (2) because one takes the evidence to show that the neural correlates of the relevant conscious processes are not "hooked up" in the right way to the neural processes that cause behavior—they occur too late or in the wrong place to get in the causal loop leading to action.

Taking the first option would make specific scientific discoveries largely irrelevant to debates about free will or mental causation. It would be motivated by, and supported by, philosophical arguments, such as the "causal exclusion argument" (Kim, 1998), which neither rely on, nor need, evidence from the mind sciences to go through. These arguments *begin* by assuming that all conscious processes (events, states, properties) correlate with neural processes (events, states, properties). Then they assume that the neural processes do all the causal work, leaving no role for the conscious properties to do any work as *distinct* properties.

Such arguments are contentious and, I believe, unsound.[12] On many theories of causation, there is no reason to say that *only* the lowest physical level of properties can do any real causal work. The fact that birds are composed of quarks does not mean that their wings play no causal role in flight. If conscious mental states are composed of neurons, that does not mean that the neurons cause (or explain) all behavior. Indeed, if these arguments work, then it is not clear how neurons could do any causal work *as neurons* since all the causal work would be done by the causal interactions among the quarks (or whatever the lowest physical level turns out to be) that compose the neurons.[13] In any case, discoveries in cognitive

science do not add much to these ongoing debates about causal exclusion and "metaphysical epiphenomenalism." If this sort of epiphenomenalism is the purported threat to free will, it is *not* because science establishes it.

Instead, specific discoveries in neuroscience and psychology usefully inform the debate only if one takes the second option in the dilemma described above. This is a position I call "modular epiphenomenalism" (Nahmias, 2002). Again, it begins with the naturalistic assumption that conscious mental processes have neural correlates, and then it suggests that *those* neural correlates are not causally relevant in producing our actions. Using the shorthand of "modules" (i.e., somewhat encapsulated cognitive systems or processes), modular epiphenomenalism claims that those modules involved in conscious decisions or intention formation do not produce our behavior; rather other modules or processes that involve no conscious states produce our behavior. The conscious processes occur too late, or in the wrong place, to cause our actions. They either get the news about what we're doing as it's happening or after the fact, and they create the illusion that conscious processes are the causal source of what we're doing (see Wegner, 2002).

Modular epiphenomenalism (at long last!) provides us with a thesis that the mind sciences can provide evidence for or against. And it provides a thesis that would, if true, raise serious concerns about free will since both ordinary intuitions and philosophical theories, compatibilist and incompatibilist alike, suggest that relevant conscious mental processes need to play some causal role in actions that we count as free and responsible. If *bypassing* is true, then we lack free will. While determinism, naturalism, and metaphysical epiphenomenalism are theses about the "form" of causation, each pitched at a level such that they are unlikely to be informed by discoveries in cognitive science, modular epiphenomenalism is a thesis about the "content" of the causal processes that lead to action, and it can be, and has been, usefully explored scientifically.

The Role of Consciousness in Action

Despite all the ground clearing so far, the philosophical analysis cannot come to an end yet since a lot depends on *which* conscious processes are relevant to free will and whether *they* are epiphenomenal. We can see this if we consider Libet's (1985) oft-discussed research, as well as more recent versions of his paradigm. Libet demonstrated that voluntary muscle movements (flexing one's wrist) are regularly preceded by "readiness potentials" (RPs), brain waves in the supplementary motor area (SMA) which occur

about half a second (500 milliseconds) before the movement. Libet also had participants report when they became aware of the "intention, desire, or urge" to move, and this measure suggested that awareness (time W) occurred only 150 milliseconds before the movement—350 milliseconds *after* the RP. Libet (1999) concluded that voluntary actions "begin in the brain unconsciously, well before the person consciously knows he wants to act" (p. 51). And he interpreted this result as evidence that our conscious intention to move is not the cause of our movement but, like the movement itself, an effect of earlier (nonconscious) brain activity.

Libet sometimes suggests that conscious intentions are nonphysical events and interprets the threat to free will in terms of naturalism or metaphysical epiphenomenalism. He wonders whether "conscious will may, at times, exert effects not in accord with known physical laws" (Libet, 1999, p. 56). And illustrating his (mis)understanding of "determinism" in terms of naturalism, he writes,

> But we have not answered the question of whether our consciously willed acts are fully determined by natural laws that govern the activities of nerve cells in the brain, or whether conscious decisions can proceed to some degree independently of natural determinism.... Quantum mechanics forces us to deal with probabilities rather than with certainties of events.... [but] they might nevertheless be in accord with natural laws and therefore determined. (p. 55)

Since quantum indeterminism is clearly inconsistent with determinism, Libet is instead suggesting that consciousness cannot play the appropriate role in action if it is understood in naturalistic terms—that is, if conscious processes have neural correlates governed by natural laws. Nonetheless, even if we reject these dualist assumptions, as I've suggested we should, Libet's data might still look like evidence for modular epiphenomenalism: The RP in the SMA is a nonconscious process that causes the movement while the neural correlates of the conscious intention to move are shown to be epiphenomenal because they occur too late to influence the movement.

Libet's paradigm has been replicated and extended in numerous other studies, including a recent study that used single-neuron recording (Fried et al., 2011) and the fMRI studies by Haynes's group described above (Soon et al., 2008), from which the authors conclude that "two specific regions in the frontal and parietal cortex of the human brain had considerable information that predicted the outcome of a motor decision the subject had not yet consciously made" (p. 545; see also Soon et al., 2013). Assuming that further studies could drive the predictive accuracy much higher

and assuming this model of agency applied to all behavior—two big assumptions—modular epiphenomenalism and premise 2E gain plausibility: Nonconscious neural processes cause actions that we experience as freely chosen, while conscious processes merely *observe* unconsciously formed decisions rather than *making* them.

However, this conclusion depends on (at least) three questionable assumptions: (1) identifying the nonconscious neural activity that precedes awareness as the "motor decision" or intention to act, (2) concluding that this neural activity *bypasses* the processes involved in conscious intention formation rather than working *through* those processes, and (3) identifying participants' reports of conscious awareness with a conscious *decision* (and identifying the time of those reports with the time of their conscious decision). If (1) or (2) is false, then the nonconscious activity measured in the experiments may simply represent causal precursors to, or activity building up to, the formation of conscious decisions or intentions rather than representing an actual *decision* that is sufficient for the movement to occur and that allows no causal role for later neural processes that underlie consciousness. If (3) is false, then participants may simply be reporting an awareness of an *urge* to move, rather than a conscious decision, having followed the experimenters' instructions *not* to plan to move or to push a particular button ahead of time but to be more passive in their actions.

The existing data do not establish any of these three assumptions. For instance, contra (1), RPs in the SMA and the even earlier activity in the frontopolar cortex may represent the brain activity underlying nonconscious *urges* to move soon (or to push the left or right button) rather than anything properly labeled *intentions* or *decisions*. On this interpretation, this nonconscious activity then *usually* causes a conscious experience reported by participants (presumably by causing the relevant neural correlates of such experiences), but in some cases the urge may be "vetoed," perhaps by participants' conscious intention not to act on that particular urge, or the action may need to be "triggered" by a conscious intention (see Mele, 2009). The data are simply unable to show that nonconscious neural activity is a sufficient cause of particular actions. Libet did not even include in his analysis cases where the participants felt the urge to flex but did not actually flex. In the Haynes study, recall that they found the early brain activity predicted the choices at only 10% above chance, so this evidence does not show that later conscious thoughts, whose neural correlates were not captured in the analysis, are causally irrelevant to which button was pressed and when. This interpretation allows that the neural

activity underlying the consciously experienced decision can still causally influence when and whether the person acts.[14] Indeed, while it is possible that future research will allow increases in predictive accuracy, it is simply impossible that any neural activity occurring 7 to 10 seconds before action predicts what people will do with 100% accuracy since we know that people can react to cues in much less time. No neural activity can guarantee a movement 7 seconds later since, *after* it occurs, participants can still react to an experimenter saying, for instance, "OK, now don't press any buttons for the next 10 seconds" (without such an ability, we'd all have died in car accidents by now!). Presumably, we can change our own minds during such time spans as well. To be clear, the issue here is not whether changing our minds (or vetoing urges) also has neural correlates—we're assuming it does—but whether the neural correlates of our conscious mental activity have any effects on what we do. If they do, then conscious processes are not (modular) epiphenomenal.

Another possibility challenging assumptions (1) and (2) is that the early nonconscious brain activity detected in these experiments does not represent a *decision* or *intention*, but instead either *is* (part of) the correlate of the conscious decision or represents part of the necessary buildup to such decisions. After all, if we assume that conscious processes have neural correlates, then we should expect that conscious experiences do not arise out of nowhere and in no time (see Dennett, 1991). Rather, they will be produced by earlier complexes of events, including external stimuli and neural activity, some of which may have been caused by even earlier conscious processes. For instance, in these experiments the participants presumably consciously processed the experimenters' instructions, which in the Haynes study were "to press either the left or right button with the index finger of the corresponding hand immediately when they became aware of the urge to do so" (2008 supplementary material, p. 15) and "to avoid any form of preplanning for choice of movement or time of execution" (p. 17) (Libet's instructions were similar). If participants followed these instructions, they formed a *distal* intention (or plan) to allow an urge to press one of the buttons to arise within them and then pay attention to when it arises. As such, it is likely, on the one hand, that this (conscious) distal intention causally influenced the spontaneous generation of nonconscious urges to act, and on the other hand, that participants are not really reporting a consciously formed *intention* or *decision* to act now but rather are reporting the time at which they felt an *urge* to act, contra the third assumption above (see Mele, 2009).[15] Because these experiments involve several dozen trials, it is even more plausible that people develop an action

plan to allow urges to move to come upon them and let those urges proceed to action.

Indeed, in a recent study using a Libet paradigm, Pockett and Purdy (2010) found differences in participants' reports of the time of awareness (W), depending on whether they were instructed to report when they experienced an *urge* to press one of two buttons or when they made a *decision* to press one of the buttons. The event-related potentials for the different trials were also different. Furthermore, many participants reported awareness of decisions before awareness of urges and before Libet's RP onset of about 500 milliseconds before movement. Trevena and Miller (2009) also present results that suggest that the RP is not a correlate of a decision to move but of *preparation* for a decision either to move or not to move.

There are a variety of other interpretations and responses to Libet's experiment and to subsequent research (see, e.g., essays in Sinnott-Armstrong & Nadel, 2011, and Klemm, 2010). Many of them develop the points I am emphasizing—that we should expect preparatory brain activity to occur prior to decisions but that the evidence so far does not show that the neural correlates of conscious decisions or intentions occur too late to influence—or occur on a sidetrack away from—the processes that most proximally control bodily movement.[16]

Nonetheless, it is still possible that the relevant evidence will come in to show that when we consciously intend an action just before we act, our being conscious (and its neural correlates) simply occurs too late to causally influence the action (or the neural correlates do not occur on the pathway to behavior control). Even *if* this turned out to be true, however, I do not think it would represent a significant challenge to free will. Consider your own experiences of most voluntary action. If they are like mine, they rarely involve specific conscious intentions to move in particular ways just prior to moving. Rather, they are preceded by more distal, and more general, intentions or plans to carry out various actions, followed by conscious monitoring of what we're doing to make sure our actions correspond to these general intentions or plans.

For instance, in these experiments, even if the proximate conscious urge or intention to move occurs too late to affect the action, it would not follow that all conscious mental states were epiphenomenal, since it has not been shown that participants' consciously agreeing to move when the urge strikes them played no role in their later actions. Similarly, when we drive or play sports or prepare meals, we do not generally form conscious intentions to perform each of the component actions of these activities.

When we lecture to students or converse with friends, we do not consciously consider exactly which words we are going to say right before saying them. Rather, we may consciously consider what *sorts* of things we want to say and then we "let ourselves go," though we consciously monitor what we say and we may stop to consider how we should proceed, for instance, in response to what our interlocutor says.[17]

On many theories of free will, what is essential is not that we have conscious intentions *just prior* to action or that our being aware of these proximal intentions produces our actions, but rather that our conscious deliberations, plans, and distal intentions (or, assuming naturalism, their neural correlates) can have proper downstream effects on how we act in the relevant situations. Such conscious causation would allow a relevant role for our deliberations among projected alternatives for action and consideration of which alternative accords with reasons that we have (at some point) consciously accepted, for our planning how to carry out complex series of actions, and for our controlling behavior in the face of conflicting desires. There is simply no evidence yet to show that such conscious deliberation, reasoning, and planning lack these causal effects on what we do or that our conscious monitoring of our behavior is not critically involved in how we carry out and adjust our actions. On the contrary, there is evidence that conscious "implementation intentions" influence actions. For instance, people are more likely to follow through on a resolution or plan when they consciously form an intention to act at a certain time than when they do not form such an intention (Gollwitzer, 1999). Furthermore, Baumeister, Masicampo, and Vohs (2011) provide other examples of behaviors that are improved by conscious reasoning and conscious attention to action (see Baumeister, this volume).[18]

To conclude this section, consider my fictional brain imaging study once again. In it students are asked to consider reasons for and against each of three psychology classes for up to a minute before picking one. If studies like this are supposed to challenge free will, it is not because they establish determinism or naturalism—they don't. If we assume naturalism, as the article suggests, then all of the students' mental activity, including their conscious deliberation, has physical (e.g., neural) correlates. The question, then, is whether *those* neural correlates play an appropriate role in the students' decisions. If they do, then their conscious mental activity is *not* bypassed. That those neural correlates have causal antecedents, even deterministic ones, does not undermine their causal role—just because an event E is caused does not show that E has no effects. It is conceivable that the neural correlates of conscious deliberation are not hooked up to the

neural processes that form intentions and produce behavior—all the students' deliberations could just be spinning wheels. However, the existing neuroscientific evidence has not established anything like this. And it would be quite surprising if all of the metabolically expensive neural activity subserving our conscious deliberation was a causal dead end—the appendix of the brain.

Rationality and Rationalization

Nonetheless, empirical evidence from neuroscience and psychology could show that the causal impact of conscious mental processes is limited, and a plausible theory of free will and responsibility must take into account such evidence. Indeed, some research suggests that, more often than we think, our actions do not accord with reasons that we have consciously considered or that we would accept were we to consider them. Research on moral judgment and behavior suggests that when people make moral judgments, they often act on immediate gut reactions and then their conscious reasoning just comes up with post hoc rationalizations for these gut reactions (e.g., Haidt, 2001; Greene, 2007). And research in social psychology suggests that we often are influenced by situational factors of which we are unaware and whose influence we would not accept were we to know about them. For instance, such research suggests that whether we help someone in need depends less on whether the person needs help or whether we consider ourselves to be helpful than on factors we do not recognize as influencing us, such as the number of bystanders, the ambient noise, or whether we are in a hurry. And these factors are not ones that people tend to accept as good reasons for failing to help.[19] Such results have been generalized to suggest that we are "rationalization machines"; psychologist Roger Shank writes, "When people try to rationally analyze potential options, their unconscious, emotional thoughts take over and make the choice for them.... Decisions are made for us by our unconscious, the conscious [mind] is in charge of making up reasons for those decisions which sound rational" (www.edge.org, 1/5/05).

This view suggests one more challenge to free will that I will call the "argument from rationalization":

1R. Free will requires that one's actions properly derive from reasons for action that one has at some point consciously considered (or at least that one would accept if one considered them).
2R. Science is showing that our actions do not properly derive from reasons that we have consciously considered or would accept as reasons for action. Rather, our actions are produced by other (nonconscious) factors and we often *rationalize* them after the fact.
3R. Thus, science is showing that humans lack free will (and moral responsibility).

Premise 1R is plausible, and many philosophical theories of free will, both compatibilist and incompatibilist, take something like it to be a necessary condition for free will. Some of the evidence for modular epiphenomenalism I discussed above might be taken to support premise 2R, but the neuroscientific research alone does not properly support it as a general truth. The moral psychology and social psychology research is more relevant since it offers evidence of cases where we don't know why we do what we do and where we make up reasons for why we did what we did. Unlike the potential threats of determinism, naturalism, or metaphysical epiphenomenalism, which are based on the "form" of behavior causation, this psychological research is at the right level to inform us about the "content" of the causal processes leading to action and the scope of our capacities for free will.

Nonetheless, this research has not established that conscious reasoning is *always* post hoc and inefficacious, and I suspect it will not establish such a sweeping conclusion. Instead, it is suggesting, and it may further show, that we have *less* free will than we tend to think we have. Hence, such scientific research challenges our *degrees* of freedom. Our free will is not unlimited. Rather, the evidence suggests limitations to the extent to which we possess the capacities required for free will and the extent to which we can exercise those capacities. And this suggests we may not be morally responsible for our actions to the extent that many assume.[20] On the other hand, such research can also provide information about how we can overcome some of these limitations, thereby increasing our freedom and responsibility (see Nahmias, 2007).

Conclusion

Let me conclude by listing some things we know and don't know about free will and related concepts, and the contributions the modern mind sciences might offer to our knowledge:

1. We don't know whether or not *determinism* is true. We do know that the sciences that study human brains and behavior are unlikely to establish whether or not universal determinism is true. We also have good reason to believe that the answer to this question about determinism is

less important to ordinary people than tradition suggests. Other potential challenges matter more.

2. We don't know whether *naturalism* is true. We don't know for sure whether there are nonphysical minds or laws of psychology that float free of the laws that govern the rest of the universe. But we have extremely good reasons to doubt such nonnaturalism, and cognitive science continues to provide inductive evidence for naturalism. However, we have every reason to believe that naturalism about free will—understood as a set of cognitive capacities that science can study—is plausible and that most people are amenable to this possibility.

3. Even if we have good reason to accept naturalism, including the idea that all conscious mental processes have neural correlates governed by laws of nature, we do not yet have a theory of consciousness that allows us to understand the relationship between the conscious mind and the brain. In my view, it is this lack of understanding that so easily leads people, including some scientists, to think that naturalism rules out free will. When we are told that neural processes *XYZ* explain certain behaviors, and we do not understand the relationship between *XYZ* and the conscious processes that precede those behaviors, then it is very easy to conclude that the conscious processes don't do anything—that they are causally epiphenomenal. I suspect that *until* cognitive science, assisted by philosophy (or vice versa), develops a naturalistic theory of consciousness, increasing information about what exactly happens in the brain when we act will look like a challenge to free will, because such information will look like it conflicts with our folk understanding of how our conscious deliberations and reasons cause behavior. In short, if free will appears to be an illusion, it is because of our ignorance about the mind–body relation. Conversely, I predict that *if and when* we have a theory of how conscious mental processes influence behavior *because* of their relationship to the relevant underlying neural processes, then we will find that the problem of free will largely dissolves.[21]

4. We also do not yet know how much of our behavior is influenced by the neural correlates of conscious mental processes. Nonetheless, cognitive science has provided no evidence that *distal* plans and conscious deliberation, or their neural correlates, are epiphenomenal. Instead, there is evidence that they do play a causal role in some of our behavior, though we do not know how much.

5. We also do not know how much of our behavior is rational, according to our own reflective judgments about what we should do. Our capacities to act in accord with reasons that we have accepted, or at least reasons we would accept, is important for autonomous and responsible agency. If future research (e.g., in social psychology) suggests that these capacities are limited, then it will thereby suggest limitations to free will and responsibility. However, it might also suggest ways for us to overcome some of these limitations.

My overall conclusion, then, is that we do have free will, though it is limited, so we need to learn how to develop it and to use it wisely.

This limited-free-will view is progressive in a certain way. Skepticism about free will—at least as it is likely to be perceived by most people—risks undermining people's belief in the capacities necessary to advocate working hard to improve one's position, to take responsibility for one's failures, to exert willpower in the face of weariness, and to deliberate carefully among alternatives to make good choices—that is, to make personal and moral progress. The limited-free-will view, on the other hand, provides room for such virtues while it also suggests increased tolerance and compassion for people unfortunate enough to lack sufficient capacities for rational self-control. This view can counter an unlimited-free-will view that some people, especially in America, seem to hold, one that suggests people *completely* deserve everything that happens to them, good or bad. Realism about the limits of free will, along with a realistic and empirically informed understanding of our capacities, is both more forgiving than an unrealistic theory of unlimited free will and more hopeful and fruitful than a skepticism that risks erasing useful distinctions between (more) free and unfree actions.

Acknowledgments

Thank you to Walter Sinnott-Armstrong, Jason Shepard, Thomas Nadelhoffer, my 2012 Moral Psychology class, and audiences at Agnes Scott College, Duke University, Georgia State University, and Bielefeld University. This chapter was completed in part with support from a BQFW grant from the John Templeton Foundation. The opinions expressed in this article are my own and do not necessarily reflect the views of the John Templeton Foundation.

Notes

1. These claims also receive a lot of discussion on widely read blogs and in science publications such as *NewScientist*, *Nature*, and *Science*. The only articles I know of that have presented significant responses to scientists' claims about the illusion of

free will are *Nature*'s (2011) "Taking Aim at Free Will" and my *New York Times* (11/13/11) article "Is Neuroscience the Death of Free Will?" Receiving less media attention are the few scientists who argue that their research helps to explain free will, rather than explaining it away (see, e.g., Baumeister, this volume, and Newsome, this volume).

2. Participants could also respond that the studies provided evidence neither for nor against these implications, which explains why percentages reported in the text do not add up to 100. Differences in responses reported in the text are statistically significant. The different presentations of the study also had a significant influence on people's responses on a new scale designed to measure strength of belief in free will and agency (Nadelhoffer, Nahmias, Ross, Shepard, & Sripada, in preparation 2013). Responses on similar scales have been shown to mediate behavioral changes in response to scientific claims about free will (e.g., Vohs & Schooler, 2008).

3. This claim would presumably entail that people do not genuinely deserve reward and punishment for their actions. We could still engage in the consequentialist practices of punishment, imprisoning criminals in order to deter them and others from crime and to rehabilitate them, but retributive justifications for punishment would be unwarranted (see Pereboom, 2001; Greene & Cohen, 2004; Harris, 2012).

4. Another incompatibilist argument focuses on the fact that determinism entails that we cannot be the "ultimate source" of our actions, because it entails that there are ultimately conditions for those actions that can be traced to conditions over which we had no control (Strawson, 1986; Pereboom, 2001). However, this argument works just as well if determinism turns out to be false because the universe includes some indeterministically caused events, since such events can equally be traced to conditions over which we had no control. If determinism is incompatible with free will or responsibility because it means we cannot be the "ultimate source" of our actions, then science adds nothing to this argument, at best making the causal story more salient to people (e.g., as suggested by Greene & Cohen, 2004).

5. This definition of determinism does not mention causation. Determinism might also be defined as the thesis that every event is completely caused by earlier events, such that the later event *had* to occur, *given* the earlier events and the laws of nature. On this definition, it remains true that sciences that study humans are simply not in a position to establish determinism. If scientists mean by "determinism" something like "causation by prior events" (whether probabilistic or deterministic), then they likely have in mind one of the challenges described in the following sections.

6. Some influential contemporary compatibilist theories include Strawson (1962), Dennett (1984), Frankfurt (1988), Wolf (1990), Fischer and Ravizza (1998), and Mele (1995) (for others, see McKenna, 2004). In a recent survey of almost 1,000 philosophy faculty, 59% identified themselves as compatibilists about free will and determinism versus 14% libertarian, 12% "no free will" (most of whom are likely

incompatibilists), and 15% "undecided" or "some other position"; the distribution was not significantly different for those who specialize in philosophy of action (see http://philpapers.org/surveys/results.pl).

7. In Nahmias, Coates, and Kvaran (2007), we found that most people responded that determinism, when framed in terms that included psychological states, such as thoughts and desires, does not rule out free will and responsibility, but when determinism was framed in terms of neural and chemical states, most people did take it to threaten free will and responsibility. This is especially relevant, given that the scientific claims about free will typically do not describe determinism per se but instead describe reductionistic causation of behavior in terms of brain states. Other work in experimental philosophy suggests that it is folk compatibilist intuitions that are the result of error, driven largely by the emotions invoked by specific cases of agents doing immoral acts (Nichols & Knobe, 2007). All existing work finds that people are more likely to ascribe free will and responsibility to concrete descriptions of specific agents, compared to abstract descriptions. While some argue that abstract cases elicit more reliable intuitions, I think that ascriptions of freedom and responsibility are more likely to be reliable when people are considering specific agents and actions, engaging our capacities to attribute relevant mental states and capacities to those agents.

8. "Naturalism" is itself a slippery concept. Here, I mean it primarily to contrast with the "nonnaturalism" that some scientists assume free will requires. My arguments in this section should carry through if we understand "naturalism" to mean "physicalism" (in the sense that everything that exists supervenes on the physical).

9. See note 6 for naturalistic compatibilist accounts, and for attempts to develop naturalistic libertarian accounts, see, for example, Kane (1996), Balaguer (2009), Clarke (2003), and arguably O'Connor (2000).

10. Data are from pilot studies for Nadelhoffer, Nahmias, Ross, Shepard, and Sripada (in preparation 2013; see also results described in Mele, this volume, Monroe & Malle, 2010, and Monroe, Malle & Dillon, under review). Such definitional questions are not ideal for understanding people's conceptual usage, and other results suggest that people's usage is more consistent with libertarian theories of free will. For instance, most people respond that we have nonphysical souls and that human action can only be understood in terms of our souls and minds and not just in terms of our brains (see Nadelhoffer, this volume). The primary conclusion drawn here is that most people do not seem inclined to reject free will or moral responsibility in light of the possibility that we will gain a naturalistic understanding of human minds and behavior (see Nahmias & Thompson, in press).

11. Even if ordinary people accepted a dualist theory that these cognitive capacities are carried out in a nonphysical soul, cognitive science would suggest that we *revise* that common view, not that we conclude that language and reasoning are illusory

or that we should stop attributing these capacities to humans. It would be bizarre for a neuroscientist (paraphrasing Montague, 2008) to claim that, because language and reason are commonly thought to "emerge wholly formed from somewhere indescribable and outside the purview of physical descriptions," language and reasoning are thereby "not even in principle within reach of scientific description." Similarly, for the minority of people who are committed to nonnaturalism about free will, the proper response is to *revise* their view, not to adopt their mistaken definitions in order to conclude that free will is an illusion (cf. Vargas, 2009).

12. There are many responses to causal exclusion arguments, offering explanations of how to understand the causal role of conscious mental states on the assumption that they supervene on physical states—for example, Wilson (2009), Bennett (2008), and Woodward (2008).

13. This debate about which levels of explanation count as doing real causal work might be seen as a debate between different disciplines within cognitive science: Do the mental states that some psychologists study and use in their theories, such as beliefs, desires, decisions, and plans, count as *real* (e.g., just as real as other theoretical entities in science), and do they make a causal difference? Or, will neuroscience be able to explain all behavior without reference to such mental states, suggesting a type of eliminativism or epiphenomenalism about them (see, e.g., Craver, 2007)?

14. This alternative account is consistent with Libet's own view that the conscious will has "veto power," but the way he describes this possibility suggests dualism, whereas this account does not.

15. When I ask students to replicate Libet's paradigm, many report experiencing the "decision" to move at the moment they move and seem surprised that it should be expected to come earlier since they were told not to plan when to move. Many also report my own phenomenology—that what I am aware of seems more like an urge or desire to move than a decision or intention to move.

16. I have not discussed the evidence presented by Wegner (2002, 2008) or Bargh (2008) for similar conclusions that the experience of conscious will is an illusion. I discuss Wegner in Nahmias (2002, 2005). In general, I take their evidence to suggest that we can sometimes be mistaken about whether our conscious intentions causally influence our actions and that nonconscious processes can significantly influence our actions (and more than we expect), but it does not support the general conclusion that conscious intentions (including distal ones), and their neural correlates, are always causally cut off from action control.

17. I take actions, such as fluent conversations, that accord with our (earlier) conscious thoughts and plans to be plausible, perhaps paradigmatic, examples of freely willed actions, ones for which we can be morally responsible. However, Bargh (2008) suggests just the opposite when he says, "Our ability to take a vague thought and have it come out of our mouths in a complete coherent sentence, the production

of which happens unconsciously, is a paramount example of this [integration of separate, parallel inputs into serial responses]. *It is not something we need consciousness or free will for*" (p. 145, his italics).

18. Indeed, Haynes's own fMRI studies (Soon et al., 2008) suggest that the area of frontopolar cortex (Brodmann area 10) that predicts participants' decisions also appears to store action plans and hold intentions between conscious formation of them and action. Hence, it might help to link our formation of distal intentions and plans with the appropriate actions.

19. For overviews of such research see Ross and Nisbett (1991). The challenge to free will from situationist research has been discussed by Nahmias (2007) and Doris (2002, chapter 7) (see also Churchland & Suhler, this volume).

20. The idea that free will can be possessed or exercised to varying degrees is unorthodox in philosophy. I believe, however, that it is plausible on both compatibilist and libertarian accounts of free will and that it accords with the way ordinary people understand free will and understand its relationship to moral responsibility, which we tend to attribute to people to varying degrees.

21. What may linger will be the worry described in note 4—that the sort of "ultimate responsibility" that requires self-creation is impossible. That worry, I believe, is one that arises largely in the context of philosophical discussions and is typically dismissed outside of those contexts, as are most skeptical theses, and rightly so, in my view, since the arguments for such skepticism rely on principles that are plausible when applied to many specific examples but should ultimately be rejected as universally applicable. In this case, the "backtracking" principle that should be rejected says something like this: For any action Y, an agent can do Y freely, and be morally responsible for Y, only if the agent was free and responsible in doing X, where X brings about the agent's doing Y (cf. van Inwagen's principle Beta, 1983, and Strawson's Basic Argument, 1986).

1.3 Response to Misirlisoy and Haggard and to Björnsson and Pereboom

Eddy Nahmias

The responses by Erman Misirlisoy and Patrick Haggard (M&H) and Gunnar Björnsson and Derk Pereboom (B&P) provide very useful ways to highlight the issues I raise in my chapter and the disagreements between competing positions in debates about free will. I thank the four of them for providing such thoughtful and challenging responses. I will begin by pointing out where we agree, and then, of course, I will point out why I think they are mistaken.[1]

My commentators and I agree that humans lack what many incompatibilists think is essential to free will and moral responsibility, including agent-causal powers or the power to be "ultimate difference makers" (as B&P define it). We also agree that scientific evidence and philosophical arguments provide convincing reasons to reject dualism and to accept physicalism or naturalism, defined loosely as the view that everything that exists, including minds, is composed only of things that physics can study and is subject to the laws of nature. This view entails, as M&H put it, that "There is no thinking 'I' independent of the brain."[2]

Finally, we all agree that we should be concerned primarily with free will understood as the set of powers or abilities required to be morally responsible—that is, potentially to deserve blame or praise, punishment or reward. This is the concept of free will that M&H tie to "a strong concept of personal and social responsibility" and that B&P define in terms of "basic desert."[3]

With these agreements laid out, we can see more clearly that we disagree about what is required for this type of free will:

1. B&P reject my view that free will does not require "ultimate difference making" and disagree that my view best accords with the ordinary understanding of free will and responsibility.

2. M&H similarly reject my understanding of free will and of ordinary intuitions about it, and they conclude that the "intuitive sense of free will cannot readily be reconciled with the available scientific evidence."

I believe my experimental evidence undermines these claims about ordinary intuitions about free will and responsibility.[4] In my response here, however, I will not focus on that evidence, but instead I will try to diagnose the sources of my commentators' views and challenge their apparent appeal. To do so, I will first pick up on B&P's useful introduction of the terms "difference maker" and "independent variable," and I will then extend the hypothetical neuroscientific study presented in my chapter to argue that our conscious reasoning can be the sort of difference maker that matters to free will.

B&P define "ultimate difference making" as requiring "that the difference maker is *an independent variable* in the causal system of the universe, that is, a variable the value of which is not determined by the value of other variables in that system." This sense of "difference making" may be, as they suggest, "perfectly legitimate," but if so, it is remarkably stringent and not very useful.

If our universe is deterministic, then their definition would entail that there are simply *no* independent variables (or at most, there could be just *one* if there is an initiating cause of the universe that still counts as a variable *in* the system). All variables in a deterministic system would be "determined by the value of other [earlier] variables in that system" since "other variables" can encompass the entire state of the universe, or whatever parts of it are causally relevant to the variable in question. Of course, incompatibilist arguments work by pointing out that determinism has precisely this consequence and by defining free will and responsibility such that they require that agents are "independent variables" in precisely this sense. (As B&P point out, indeterminism would not allow any variables to influence the objective probabilities of what happens based on preceding events, so the causal powers of indeterministic variables would also be ultimately "determined by the value of other variables.")

Because these notions of ultimate difference making and independent variables are so stringent, they are also essentially useless. They cannot help us individuate variables or discern or dispute which things, events, or processes ultimately make a difference in the real world, in our interactions with each other, or in our scientific explanations.[5] For instance, since neuroscientists cannot know whether determinism is true, they are not in a position to discern, according to B&P's definitions, whether the firing of

a specific group of neurons (e.g., in motor cortex) is an independent variable that ultimately makes a difference to an organism's behavior. Their definition also makes specific discoveries in neuroscience and other sciences irrelevant to understanding whether we have free will since such discoveries cannot tell us whether or not we are "ultimate difference makers" or whether our mental states count as "independent variables" (points I emphasized in a different guise in my chapter). If we use B&P's definitions, then M&H's worthy goal of building a "neuroscientific database" to understand free will and responsibility becomes largely irrelevant.

People certainly understand agents and choices differently than other causal variables—most notably, we consider the role that human agents' reasons, goals, and intentions play in their choices and actions. However, it is implausible that most people have metaphysical presuppositions entailing that the distinctions we make regarding agency and responsibility would be eradicated by determinism, in part because we cannot discern whether or not determinism is true. B&P raise concerns with my studies on ordinary people's intuitions, but they do not address the fact that the majority of participants across conditions do *not* make the mistake of conflating determinism with bypassing and do *not* take determinism to rule out free will or responsibility. If B&P (or scientists) want to argue that free will, as ordinarily understood, requires *ultimate* difference making (or the falsity of determinism), then they have to argue that most participants in these studies are making a mistake or failing to understand the implications of determinism.[6]

For these and other reasons, I suggest a slightly altered definition of ultimate difference making and independent variables:

An ultimate difference maker is an independent variable in the causal system of the universe, that is, a variable which is not determined by the value of *any other variable* in that system. (My italicized words replace B&P's "other variables" and thus my definition allows independent variables to exist in a deterministic universe.[7])

The intuitive idea is that an ultimate difference maker D is a causal variable that is the locus of many causal inputs such that none of these inputs (i.e., *any other variable*) can be picked out as the cause of D's effects. No other variable determines D's effects. Hence, for most purposes, nothing explains D's effects as well as D itself. On the "metaphysical side" difference makers are "causal funnels," the source of their effects by integrating a range of earlier causal influences. Because of this, on the

"epistemic side" they are explanatorily useful and ineliminable—they are the variables we use to explain causal relations and to draw distinctions about what causes what. They are also the variables on which people (e.g., scientists) intervene to make differences to what happens in the world.[8] Though my definitions here have been left as rough-hewn as B&P's, they accord well with our scientific and ordinary explanatory practices of looking for variables that make a difference because of their ineliminable role in causal chains.

I will now argue that my definitions also allow us to diagnose some of the intuitions driving debates about free will: They help to explain why some neuroscientists think their discoveries challenge free will, why some people misinterpret determinism to mean bypassing, and why our intuitions about manipulation differ from our intuitions about determinism. To begin, let's consider an elaboration of the imaginary imaging study introduced in my chapter.

Imagine a Lesion Study

Given the metaphysics of mind my commentators and I agree on, we should all accept that neuroscientists could, in principle, discover the neural correlates of complex decision making—for instance, the neural activity that subserves episodes of conscious deliberation by students' considering which psychology class to take, as in the hypothetical study described in my chapter. Suppose that these future neuroscientists discover the neural correlates of such conscious reasoning (the NCs of CR). Suppose further that they are able to use some futuristic technology along the lines of transcranial magnetic stimulation (TMS) to temporarily knock out (or "lesion") the NCs of CR, and they do so during a variety of tasks carried out by a psychology student named Eve.[9]

Will Eve's behavior change while undergoing this procedure? It's hard to imagine it wouldn't. After all, how could neuroscientists have discovered the NCs of CR (in general and in Eve)? Presumably, their discoveries followed the normal procedure of correlating behavioral changes (including verbal reports) with changes in neural activity.

However, if knocking out the NCs of CR changed people's behavior only minimally (e.g., changed only verbal reports after behavior), then we'd have good evidence for what I labeled *modular epiphenomenalism*, and not just the sort potentially suggested by existing neuroscientific studies on conscious *proximal* intentions. Rather, this would be *massive* modular epiphenomenalism vindicating what I called the argument from rationalization: None of the neural activity associated with conscious deliberation, decision making, planning, or conscious *distal* intentions significantly influences downstream behavior. That result would indeed provide evidence that we lack a capacity for conscious control over our behavior that is essential for free will and responsibility. Our conscious reasoning would *not* be a difference maker.

Existing evidence suggests that such an extreme result is unlikely (see, e.g., Baumeister, this volume), though I concluded my chapter by pointing out that the degree to which our behavior is influenced by our reasoning (or by reasons we would endorse) is an open empirical question. This sort of scientific evidence can inform us about the degree to which we are autonomous and responsible (see Nahmias, 2007). In any case, M&H do not suggest that neuroscience will demonstrate that knocking out the NCs of CR will make no difference to behavior.[10]

Suppose then that temporarily lesioning the NCs of CR *does* change Eve's behavior in significant ways. For instance, compared to her behavior when her NCs of CR are in working order, Eve fails to make some decisions or even to act at all, or she acts very differently (e.g., in ways we would call irrational), or she makes choices in ways we can discern are abnormal (e.g., more randomly and less reasonably). Such results would suggest that the NCs of CR have important causal influences on behavior, and *massive* modular epiphenomenalism would be false.[11]

M&H write, "Nahmias argues that if our conscious intentions can have any causal effect on our actions, this is sufficient to make them free." In fact, that is *not* what I argued in my chapter. First, I argued that the efficacy of *proximal* conscious intentions may not be essential for free action, and then I argued that the causal efficacy of conscious reasoning is *necessary* for free will and responsibility. But I did not, and would not, argue that it is *sufficient*. For instance, if the neuroscientists directly stimulated Eve's NCs of CR to cause her to make particular choices they wanted her to make, then she would lack free will, as do the patients described by M&H when their SMA was directly stimulated. Or if Eve consciously reasoned about how to get a drug she was addicted to, I think that her taking the drug would, at a minimum, be *less* free than it would be had she been able to control her actions in accord with her reflectively endorsed goal to stop taking the drug.

M&H, however, suggest that, regardless of whether the NCs of CR play important causal roles in our decisions and actions (i.e., knocking them out would make big differences to our behavior), neuroscience still poses a threat to free will. Why?

I read them as suggesting two related possibilities, both of which permeate neuroscientific discussions of free will and drive the intuition that consciousness is not an independent variable or difference maker:

1. Conscious processes are not *really* causal difference makers because they have prior causes that are not conscious—that is, they are not independent variables because there is another variable (prior neural causes unrelated to consciousness) that determines their value.
2. Conscious processes are not *really* causal simply *because* they have neural correlates—that is, they are not independent variables because their value is determined by another variable (their NCs).

As I tried to show in my chapter, both of these moves are based on poor reasoning or on contentious philosophical arguments, and either way they do not allow neuroscientific evidence to inform debates about free will.

Consider (1). M&H write, "Conscious intention is certainly part of some of our actions. However, the actual cause of action is not the conscious intention but the brain activity preceding the action," and "conscious intention must be a consequence of brain activity and not a cause." I pointed out in my chapter that causes can be caused, so the fact that conscious intentions (and their neural correlates) are caused by prior brain activity is consistent with their being an "actual cause of action."

However, putting aside that mistake, perhaps the idea is that conscious intention and reasoning are not really *difference makers*. When neuroscientists describe conscious proximal intentions to flex or to push a button, it is easy to think in terms of domino-like causal chains, where the fourth domino, even if it causes the fifth to fall, is not really a difference maker since it is a variable determined by another variable—that is, the earlier dominoes' falling. Similarly, immediately prior neural activity (e.g., an RP) might determine the value—and hence the causal powers—of a conscious proximal intention (and its NCs), in which case that intention is not an independent variable, even if it is a causal link in the chain (i.e., a "part of some of our actions").

Even if this picture works for proximal intentions—and it may not—it is unlikely to work for the NCs of CR.[12] The activity carried out by the NCs of CR is *not* like a series of dominoes simply "transferring causal impetus." Like other complex neural processes, this activity involves integration and transformation of information. What happens in the NCs of CR makes a difference to what specific outputs (e.g., decisions) are produced, just as what happens in the NCs of movement preparation makes a difference to what specific outputs (e.g., arm movements) are produced. Assuming that

the NCs of CR take as input the information from the NCs of beliefs, desires, reasons, perceptions, and so forth, then those mental states will be causally relevant to one's choice, though none will individually cause that choice.

When we consider complex issues, such as how to plan our day to get numerous tasks done or which job candidate to hire, our NCs of CR are likely to be very active. There is typically no *other* variable that determines what our NCs of CR produce as output. No *individual* thoughts or desires, conscious or nonconscious, nor their NCs, determine what plan we make or which candidate we hire, at least assuming that the process required conscious consideration and integration of numerous thoughts, desires, and goals. Their integration in episodes of conscious reasoning ultimately makes a difference to decisions and behavior because no other causal variable determines the value of the NCs of CR.[13]

Of course, none of this entails that there are not large and complex *sets* of prior events that determine the particular neural activity that occurs when we consciously reason about what to do. For every specific, incredibly complex set of processes that occur in the NCs of CR on a given episode of human reasoning, there may be an incredibly complex set of prior causes that is sufficient for them to occur. Even so, there will typically not be "any other variable in the system" that determines the output (e.g., intentions, decisions, plans) of activity in the NCs of CR. If determinism is true, the set of causally sufficient conditions for episodes of activity of the NCs of CR will typically be an unwieldy and large set of events extending backward in time (and outward in the light cone of prior states of the universe). It is not true, as H&P suggest, that "[d]irectly predictive unconscious neural activity must necessarily precede any conscious intention." Perhaps unconscious neural activity directly precedes and predicts many conscious *proximal* intentions. But even so, earlier activity (the NCs of CR) may have been essential to forming those intentions (e.g., subjects' agreeing to carry out the Libet experiment). If so, it is false that "conscious intention must be a consequence of brain activity and not a cause."

Again, if the general truth of determinism is what is driving some scientists to conclude that free will is an illusion, then the specific discoveries of neuroscience do not advance the debate. We are simply led back to the standard philosophical debate, where if any science is relevant, it will be physics. And if these scientists are driven by the idea that specific *nonconscious* neural activity determines every decision we make while *bypassing* neural activity involved in conscious reasoning, then they

should predict that my hypothetical lesioning study would have no effect on behavior. I predict that the experiment would instead leave Eve without free will.

Diagnosing Bypassing

Of course, neuroscientists do not say that the complex neural activity involved in, say, visual perception or movement preparation is not an "actual cause" simply because it has (sufficient) prior causes. Rather, such activity is treated as an independent variable because there are no other individual variables that determine its values or that serve as well in neuroscientific theories and causal explanations. I'm suggesting that neuroscientists should treat the NCs of CR with the same (causal) respect.

But perhaps the problem is that they are treating the *NCs* of CR with so much respect that they assume that *consciousness* is causally irrelevant. That is, they are thinking in terms of point (2) above: that it's the neural correlates (NCs) that do all the causal work such that consciousness itself (CR) does not do any. M&H suggest this point when they write, "Neuroscience has rejected the idea that conscious intention, qua consciousness, plays any causal role in action."

In what sense has neuroscience rejected the idea that conscious processes, "qua consciousness," play no causal role in action? First, it has rejected substance dualism. Thus, if consciousness is assumed to occur in a nonphysical mind, and physical (neural) processes were shown to cause all behavior, then consciousness would play no causal role in action (as I pointed out, Libet and others sometimes suggest this alleged worry). Like M&H, I am putting aside this implausible view. Indeed, they say they are trying to prevent a problem they properly diagnose in these discussions—that "a form of dualism often seems to creep in through a side door." However, I fear they are propping open the side door themselves by assuming that consciousness is a special kind of high-level process or entity.

If instead we treat conscious mental processes like other high-level entities composed of lower-level entities—for instance, ocean waves, organisms, or neurons—then the fact that consciousness is realized by (or supervenes on) neural correlates does not thereby eliminate the causal powers of conscious mental processes. Just as waves, organisms, and neurons can be independent variables that really make a difference to what happens, so can conscious mental activity. M&H presumably think that neuroscience does, and should, reject the idea that consciousness is an emergent property with causal powers *over and above* those of its neural realizers. However, we can reject this sort of emergentism or property dualism without concluding that consciousness is causally irrelevant.

Debates about the causal efficacy of high-level properties, events, or states, including conscious mental ones and special-science ones, are contentious.[14] Many philosophers reject the metaphysical arguments against high-level causation, including B&P, who write, "we agree with Nahmias that higher-level causation is defensible." Again, these debates about mental causation are not directly advanced by neuroscientific studies. If substance dualism is false, then scientific studies simply cannot separate the causal role of the *NCs* of CR from "consciousness, qua consciousness," so discovering that brain processes (NCs) cause behavior does not, by itself, show that consciousness is causally irrelevant (see Woodward, 2008). On the contrary, if my hypothetical study showed that knocking out the NCs of CR changed people's behavior, then that should count as evidence that consciousness plays a causal role in behavior.[15]

Moreover, my schema for ultimate difference making gives us one way of understanding *why* "consciousness, qua consciousness" can be such a difference maker. Suppose many high-level psychological processes are multiply realizable in at least this sense: In the same individual, the same process can be instantiated by slightly differing patterns of neural activity. For instance, a monkey who has learned to push a button when perceiving a red target will presumably instantiate the perception of red targets with a variety of interrelated neural activations (e.g., in area V1) and will instantiate a motor preparation (or intention) to push the button with a variety of interrelated neural activations (e.g., in SMA). Neuroscientists typically do not think that the same perceptual or motor state must be instantiated by the *exact* same set of neurons in the *exact* same activation patterns, nor do they (typically) try to study neural activity in such a fine-grained way.

If so, then there are *no other variables* (in terms of particular neural realizers) that *determine the values of* these psychological variables—for example, perceptions and intentions of the monkey. On my schema, that means the psychological variables are independent variables and difference makers. This captures the idea that it is *those* high-level variables that will often best predict and explain the monkey's behavior, and it accords with the scientific practice of manipulating those variables to study both behavior and the relevant neural correlates (see Woodward, 2008). The psychological processes serve as causal funnels that best explain the behavior of the organism.

Conscious reasoning is *much* more complex than monkey (or human) button pressing. The NCs of CR are that much more likely to be multiply

realized within (and between) individuals. If so, then it is *conscious mental processes* that will often be the independent variables in human behavior rather than their neural correlates. Thus, we should reject (2)—the idea that conscious processes are not *really* causal simply because they have neural correlates.

Nonetheless, given that neuroscience studies the most complex system in the universe (the human brain) and is in its infancy, the bypassing intuition is quite understandable. We lack a scientific theory (or a well-worked-out metaphysical theory) of *how* neural processes realize conscious processes. Yet most neuroscientists (and philosophers) assume that neural processes are causally sufficient for all human behavior. It is therefore understandable that they might then have the intuition that consciousness is causally bypassed. Before Galileo's theory of inertia, the Copernican claim that the earth moves around the sun was baffling—it could not make sense of our experience of being unmoving. Similarly, without a naturalistic theory of consciousness, we can't understand how the electrochemical activity in our mushy brains *explains* our thoughts and experiences. Yet we have very good reason to adopt such naturalism (just as there were good reasons to accept the Copernican theory), and we increasingly understand the causes of behavior in terms of neural processes.

Whether or not ordinary people *explicitly* think of conscious processes as being *non*physical, they have not been offered a theory to explain how they *are* physically instantiated. As such, when told that physical (e.g., neural) processes completely determine behavior, they are likely to interpret consciousness as being bypassed. If we're told that our brains explain (cause) everything we do, without being told how our brains explain consciousness, it is easy to conclude that consciousness explains (causes) nothing we do. And it is this bypassing intuition that best explains why some people interpret both physicalism and determinism as threatening free will and moral responsibility (see Nahmias et al., 2007; Murray & Nahmias, 2012).

On the other hand, when people are simply presented with the idea that there exist neural correlates of conscious reasoning (as described in my hypothetical study and as I've used in scenarios describing nonreductive forms of determinism), they do *not* assume that conscious mental states are bypassed by their neural correlates. In that case, most people seem to accept the possibility that conscious processes happen to be instantiated in the brain *and* that they can be real difference makers.[16]

Neuroscientists like M&H who suggest that their research challenges free will face a dilemma:

Either (a) neuroscientific discoveries about human decision making present *new* empirical challenges to free will (such as massive modular epiphenomenalism),

or (b) neuroscientific discoveries simply illustrate (perhaps making more salient) *old* philosophical arguments for incompatibilism or for (metaphysical) epiphenomenalism.

I have argued that when neuroscientists, including M&H, argue for (a), they typically end up slipping into (b). Their discoveries don't add much to these philosophical arguments, and the arguments themselves are controversial. And I have argued that if we want to stay focused on (a)—if we want to develop M&H's suggested "neuroscientific database"—then we will need to explore the neural correlates of conscious reasoning and their causal role in action. Indeed, we will need to develop a neuroscientific theory of consciousness and of reasoning—no easy tasks!—but ones that have to be carried out before we can conclude that we lack free will because our conscious reasoning does not play the right role in action. (I emphasize again that I think the evidence already suggests that we possess *less* free will than we think because of limitations on the role of our rational decision making and self-control.)

If M&H are ultimately worried about the same sort of problems that B&P suggest, then my responses here may seem unsatisfying. But if so, then they and other neuroscientists should not be suggesting that they are *discovering* new challenges to free will.

Manipulation Arguments and Making a Difference

Indeed, incompatibilist philosophers are unlikely to be convinced by my claims that compatibilism is viable or that most people do not find it counterintuitive. Luckily, my primary goal in my chapter and this response is not to take on this timeless task. My goal is to show that neuroscientific results are not showing free will is an illusion for the reasons typically presented. B&P seem to accept that I've advanced that goal (see pp. 29–30). But I will briefly suggest that they should follow me further down the compatibilist path.

As mentioned above, B&P have no explanation for the many, typically majority, responses by participants in my studies that indicate no commitment to incompatibilism. B&P would need to offer an error theory explaining away these majority responses. They would also need to explain why most people do *not* make what I call the bypassing mistake and fail to

respond as they predict one should if thinking in terms of their "ultimate difference making."[17]

B&P might argue that most people do not properly understand the implications of determinism—namely, that it rules out ultimate difference making. They suggest that one effective way to illuminate these implications is manipulation arguments (see Pereboom, 2001), and others have offered "design arguments" (Mele, 2006) that share a similar structure. These arguments describe an agent (e.g., Plum) who carries out an action with outcome *O* while satisfying a full complement of compatibilist conditions but who is deterministically caused to bring about *O* by a manipulator (or designer) who ensures that Plum will do so.

Pereboom argues that there is no principled difference between Plum and an agent (say, Blum) who acts to bring about *O* in the same way in a deterministic universe. He argues that our intuition that a manipulated agent such as Plum lacks free will and responsibility is best explained by the fact that his action "results from a deterministic causal process that traces back to factors beyond his control" (2001, p. 116). If so, he argues, this should help us see that that determinism rules out free will and responsibility for the same reason.

I disagree. The best explanation for our *intuitions* about manipulated (and designed) agents does not depend on features they share with deterministic agents (see also Sripada, 2012). One such difference between Plum and Blum is that Plum (and his conscious reasoning) is *not* an independent variable or difference maker on my definition of the terms. There is another variable in the system—that is, his manipulator (or designer) and *her* conscious reasoning—that fully determines (and explains) outcome *O*. Furthermore, that variable is an agent with intentions that ensure that *O* will occur, so there is a better target of our responsibility attributions. In a deterministic universe, there is no agent other than Blum that ensures *O* will occur and no individual variable other than Blum's conscious reasoning that determines *O*. Blum and his conscious reasoning is thus the ultimate difference maker, on my definition of that term, which is not true of Plum and his conscious reasoning.

B&P might reject my interpretation of people's intuitions about such cases. If so, then we might need to run controlled studies to try to tease out exactly which factors of the cases are influencing people's judgments. I find it unlikely that people (even philosophers!) are in a position to know by introspection which features lead them to have intuitions about Plum's or Blum's freedom and responsibility. On the other hand, if B&P suggest that manipulation arguments need *not* rely on people's intuitions about the agents or why they have or lack free will and responsibility, then the arguments are unhelpful in illuminating a particular feature of determinism that should suggest we accept incompatibilism or in showing that people who offer compatibilist intuitions are making some sort of mistake.

In this response I have tried to offer further diagnoses of why some people may have the intuition that neuroscience, naturalism, or determinism threaten free will and moral responsibility. These diagnoses are meant to deflate those intuitions. Conscious processes have neural correlates, but that does not mean that our conscious minds don't matter or that our brains *make* us do what we do. And determinism would not show that our conscious reasoning makes no difference to what we do.

Acknowledgments

I appreciate helpful comments on this response from Walter Sinnott-Armstrong, Gunnar Björnsson, and Andrea Scarantino.

Notes

1. For ease of exposition, in what follows I generally refer to "B&P" but it should be noted that Björnsson and Pereboom have importantly different views about free will and responsibility, such that in some cases Pereboom and I may disagree more than Björnsson and I do. In fact, Björnsson and Persson (2012b) offer an account of people's judgments of responsibility that shares some features with my own and that rejects incompatibilism.

2. Depending on how embodied or extended the mind is, the physicalist might say that there is no thinking "I" independent of the brain *and* relevant parts of the body and the world.

3. While I agree that a central concern in the debate is the sort of free will required for basic desert in the narrow sense B&P outline, that is not the only way philosophers, scientists, or ordinary people understand "free will" or the only concern they have. Thus, even *if* basic desert were impossible, it would be too hasty to conclude that free will is an illusion. B&P recognize this, but scientists who argue against free will typically mean more than just basic desert. Some compatibilists accept that we can truly deserve praise and blame but may reject that our having free will also justifies *retributive* punishment, since there may be independent reasons for rejecting such retributivism. Other compatibilists, of a more consequentialist bent, explicitly reject this notion of ("backwards-looking") desert and punishment but nevertheless believe that free will can exist and can justify reactive attitudes and the attribution of blame.

4. For recent work, see Nahmias (2011b) and Murray and Nahmias (2012). These papers also discuss the relevance of understanding laypersons' views about free will to the philosophical and scientific debates.

5. B&P's definitions are also difficult to reconcile with most theories of causation. No theory of causation requires that something counts as an independent causal variable only if it is *not* determined by other variables. Even the libertarian notion of agent causation is arguably consistent with determinism (see, e.g., Nelkin, 2011, chapter 4).

6. To do so, they might use manipulation arguments, which I will address below.

7. If one thinks that a deterministic universe might be initiated by a variable *within* the system itself, then I would refine my definition for such a universe to say, "any other, *noninitial* variable." The main point is that once a universe (even deterministic) is "in motion," it is unhelpful to consider the *entire* state of that universe as a variable, causal or otherwise.

8. My definitions should fit with many theories of causation, but especially interventionist accounts such as Woodward (2003).

9. I hope the possibility of this thought experiment does not depend on any assumptions that my respondents would reject for reasons relevant to our philosophical disputes. Whether the experiment is feasible is a different question, and I suspect it would not be. For instance, complex conscious reasoning is presumably distributed in such a way that the imagined future technology would have to be much more precise than TMS in its ability to "knock out" specific activity. And there will be different neural correlates for different sorts of conscious reasoning.

10. M&H do say that "distal intentions are arguably less critical to the concept of responsibility.... In the legal and moral sphere, we care more about proximal urges and actual actions than we do about these distal intentions." Though we do care about "actual actions," that is not evidence that we care more about proximal urges than distal intentions. We (and the law) care deeply about whether people's actions accord with their reasons and plans, as well as their character traits and consciously endorsed values and desires.

11. Suppose Eve made the complex decision to participate in Libet's or Haynes's study and consciously planned how to carry out her "random" wrist flexes or button presses, tasks which I assume require the NCs of CR. If we *then* "lesioned" her NCs of CR, could she carry out the simple tasks of wrist flipping and button pressing, perhaps while continuing to experience the urge or intention to move? If so, that might indicate that these tasks are not ideal tests of the causal role of the conscious reasoning most essential for free will.

12. New technology allows people to control prosthetic limbs with their thoughts. How? Presumably, the NCs of the conscious proximal intentions cause the appropriate motor cortex activity, which is then interpreted by the computer to move the prosthetic limb accordingly. It seems highly unlikely that this technology could get off the ground without people forming *conscious* proximal intentions.

13. This integration of diverse and complex information in some episodes of conscious reasoning also suggests that predicting the output (decisions) would be unfeasible if not impossible, even with information about preceding neural activity.

14. See references in note 12 of my chapter. See also Pereboom (2002).

15. Note that if one accepts the logical possibility of "zombie worlds"—worlds physically identical to ours, including physical duplicates of us, but with no conscious mental states—(and I do not), then all the neuroscience experiments in such a zombie world look *exactly* the same as ours. Hence, such experiments cannot provide any information about the metaphysical relationship between consciousness and its neural correlates or about the causal properties of "consciousness, qua consciousness."

16. In current work, my collaborators and I are presenting scenarios that describe future neuroimaging technology that allows complete prediction of decisions and behavior based on specific earlier brain activity. The vast majority of participants respond that such technology is possible and that, if it were actual, people would still have free will, make choices, be responsible, and deserve blame for bad actions (see Nahmias & Thompson, in press). Such results directly contradict what some neuroscientists, such as Sam Harris (2012), assume people would say about such scenarios, belying their assumptions about the folk understanding of free will.

17. While B&P suggest that Björnsson's study provides evidence that most people interpret determinism as consistent with "throughpassing," the mean response to that question is near the midpoint of the scale (4.16). And the fact that it does not correlate with free will responses provides no evidence that most people have incompatibilist intuitions or that many, if not most, incompatibilist responses are *not* driven by bypassing judgments, as demonstrated in Murray and Nahmias (2012), especially our study 2 where bypassing was directly manipulated in the scenarios, in turn affecting responses to questions about free will and responsibility.