9

The Mind, the Brain, and the Law

THOMAS NADELHOFFER, DENA GROMET, GEOFFREY GOODWIN, EDDY NAHMIAS, CHANDRA SRIPADA, AND WALTER SINNOTT-ARMSTRONG

SETTING THE STAGE

Although there is a long-standing debate among philosophers and legal scholars concerning the nature, limits, and legal relevance of free will,[1] judges, journalists, and the public more generally view free will as important to the concept of legal responsibility. For instance, in *Morissette v. United States*, Justice Jackson claimed that "a belief in freedom of the human will" is "universal and persistent in mature systems of law."[2] Similarly, in *United States v. Grayson*, Chief Justice Burger suggested that the adoption of "a deterministic view of human conduct" would be "inconsistent with the underlying precepts of our criminal justice system."[3]

One area of the criminal law in which free will seems especially relevant is mental health law. After all, one explanation for why mental illnesses are sometimes mitigating or even exculpating is that they undermine offenders' free will and hence minimize their responsibility. As one federal judge observed in this

1. Indeed, one legal commentator goes so far as to suggest that "enough has been written from a philosophical perspective on the relationship between free will and the law that it is not easy to justify yet another such undertaking" (Green 1995, 1915).

2. *Morissette v. United States*, 342 U.S. 246, 250 (1952).

3. *United States v. Grayson*, 438 U.S. 41, 52 (1978).

context, "the concept of lack of 'free will' is both the root origin of the insanity defense and the line of its growth."[4] On this view, it is precisely because the mentally ill are sometimes viewed as having less free will that we do not hold them as responsible for their behavior.

Because free will is sometimes assumed to have this foundational role in criminal law, it is unsurprising that recent developments in neuroscience that purportedly challenge free will have generated such interest and controversy. According to some researchers, as neuroscientists uncover the neural mechanisms that undergird both normal and abnormal human behavior, we will see a radical shift in how we think about agency and responsibility (e.g., Greene & Cohen 2004). According to others, although neuroscience will continue to shed new light on how the human mind works, it will likely leave our traditional views and practices largely intact (e.g., Morse 2008).

When exploring this debate, it is crucial to make a distinction between the *descriptive* question of whether advances in the modern mind sciences will in fact change people's views and attitudes about agency and responsibility and the *normative* question of whether these discoveries *should* have a transformative effect. For present purposes, we limit our attention primarily to the descriptive question. More specifically, our goal in this chapter is to explore the potential influence that advances in neuroscience may have on legal decision makers by describing recent studies that probe folk intuitions concerning the relationship between neuroscience, agency, responsibility, and mental illness. Addressing whether recent and future advances in neuroscience *should* influence our moral and legal beliefs and practices is a task for another day.

In examining the descriptive question, we will first familiarize the reader with some of the early research in experimental philosophy on people's intuitions about agency and responsibility (section 1). We will then focus on a more specific issue—namely, whether people respond to explanations of human behavior framed in neuroscientific terms differently than they respond to explanations framed in more traditional folk psychological terms. Some parties to this debate have provided evidence that people's intuitions about agency and responsibility are differentially influenced by neural explanations than mental explanations (see Nahmias et al. 2007), whereas others have provided evidence to the contrary (see De Brigard et al. 2009). We will present the results of some new studies, which provide evidence for the former view (section 2). As we will see, explanations of criminal behavior that are couched in neural terms appear to make people less punitive than explanations couched in mental terms, especially in the context of mental illness. We will then offer what we take to be the best explanation of these differences in people's intuitions: when people are

4. *United States v. Brawner*, 471 F.2d 969, 986 (D.C. Cir. 1972) (en banc) (footnote omitted).

presented with neural explanations of human behavior, they tend to think that the agents' "deep self" (the values and beliefs they identify with) is somehow left out of the causal loop or bypassed, which in turn mitigates the agent's responsibility (section 3). In short, it is bypassing of the deep self, and not determinism per se, that seems to be motivating people's concerns when it comes to the relationships between neuroscience, agency, and responsibility. Although we provide some preliminary empirical and philosophical support for this position, more research is required to improve our understanding of people's complex and sometimes puzzling beliefs about the mind, the brain, and the law.

1. EXPERIMENTAL PHILOSOPHY, FREE WILL, AND MORAL RESPONSIBILITY

Experimental philosophy is a recent movement whose participants use the methods of psychology to probe the way people make judgments that bear on debates in philosophy. Although the movement has a name, it includes a variety of projects driven by different interests, assumptions, and goals.[5] Just in the past few years, philosophers have carried out experimental work in areas as diverse as epistemology, action theory, the philosophy of language, ethics, the philosophy of law, the philosophy of mind, and the philosophy of science. All of this work shares a two-fold commitment to using controlled and systematic studies to explore people's intuitions and to examining how the results of such experiments bear on traditional philosophical debates. In this paper, we are going to limit our attention to work in experimental philosophy on agency and responsibility. But first we set the stage with a brief discussion of the free will debate more generally.

The dominant issue in the traditional philosophical debates about free will has been whether free will and moral responsibility are compatible with determinism—that is, the metaphysical thesis that given the actual past and the laws of nature, there is only one possible future (see Van Inwagen 1983).[6] Incompatibilists, who claim that free will and determinism cannot coexist, run the gamut from pro–free will libertarians who deny the truth of determinism and suggest that we are unmoved movers (Chisholm 2003) to free will skeptics who claim that we can't be free and responsible regardless of the truth of determinism (Strawson 1986). A number of incompatibilists lie on a continuum

5. For an overview of the field of experimental philosophy, see Nadelhoffer & Nahmias 2007.

6. For helpful introductions to the major views in the free will debate, see Fischer et al. 2007; Kane 2011; and Watson 2003.

between these two extremes. The two main categories of pro–free will incompatibilist views are event-causal libertarianism (Ekstrom 2000; Kane 1996) and agent-causal libertarianism (Clarke 2003; O'Connor 2000)—each of which maintains that determinism is false and that human beings are (sometimes) free and morally responsible. Skepticism about free will and moral responsibility comes in several varieties as well (see Double 1991; Honderich 1998; Pereboom 2001; Smilansky 2000)—some of which are driven by worries about determinism and some of which are not.

There are just as many varieties of views that take free will and determinism to be compatible. Most of the original compatibilists, such as Hobbes and Hume, were known as soft determinists, and claimed that free will and responsibility actually *require* determinism (Ayer 2003; Stace 1960). Most contemporary compatibilists, however, are merely committed to the conditional view that we could be free and responsible even if the universe were deterministic. These compatibilists offer various analyses of what is required to be free and responsible agents, emphasizing, for instance, our identification with some of our desires over others (Frankfurt 1971), our ability to understand what is true and good (Wolf 1990), our sometimes being appropriate targets of reactive attitudes such as indignation or approbation (Strawson 2003), or our capacity to be appropriately responsive to reasons (Fischer 1994; Fischer & Ravizza 1998). In general, compatibilists argue that free will and moral responsibility do not require the unconditional ability to do otherwise, holding fixed the actual past and laws. Instead, compatibilists argue that free and responsible agency requires the capacities involved in self-reflection, practical deliberation, and self-control (see also Mele 1996).

Traditionally, both compatibilists and incompatibilists have assumed that their own respective views enjoy wide-scale intuitive support among nonphilosophers. Robert Kane, for instance, writes, "most ordinary people start out as natural incompatibilists... Ordinary persons have to be talked out of this natural incompatibilism by the clever arguments of philosophers" (1999, 218). Similarly, Galen Strawson argues that the incompatibilist's libertarian conception of free will, though impossible to satisfy, is precisely "the kind of freedom that most people ordinarily and unreflectively suppose themselves to possess" (1986, 30). Compatibilists also frequently appeal to commonsense intuitions, suggesting that the folk do *not* demand the libertarian requirements for free will, such as an unconditional ability to do otherwise. For instance, Daniel Dennett claims that when ordinary people assign moral responsibility, "it simply does not matter at all... whether the agent in question could have done otherwise in the circumstances" (1984, 558). William Lycan similarly argues that compatibilism is "the default position... not only true, but the only position rationally available to impartial observers" (2003, 107).

Motivated by the dearth of empirical data on what people actually think about the relationships between free will, responsibility, and determinism, Eddy Nahmias, Stephen Morris, Thomas Nadelhoffer, and Jason Turner (2005; 2006) developed some of the first studies in experimental philosophy to explore the relevant folk intuitions. Using three different descriptions of determinism, they found that a significant majority of participants (typically 65% to 85%) judged that agents in a deterministic scenario act of their own free will and are morally responsible. These early findings suggested that contrary to what incompatibilists have traditionally assumed, most people do not have intuitions that support incompatibilism. However, as is often the case when it comes to the free will debate, it soon became clear that things were more complicated than they initially appeared.

For instance, Shaun Nichols and Joshua Knobe (2007) designed and ran some follow-up studies to explore the psychological mechanisms that generate intuitions about moral responsibility. Participants were randomly assigned to either an "abstract" condition that describes a deterministic universe (A) and indeterministic universe (B) or a "concrete" condition that describes these universes but also describes a person in universe A, Bill, who murders his wife and family to be with his secretary. Whereas 72% of subjects gave the *compatibilist* response that Bill *is* "fully morally responsible for killing his wife and family" in the concrete condition, in the abstract condition 84% gave the purportedly *incompatibilist* response that it is *not* possible in universe A "for a person to be fully morally responsible for her actions."

On the surface, at least, these findings appear to put pressure on the claim that people's intuitions are robustly compatibilist. Instead, whether people are inclined to give compatibilist answers may depend less on the presence (or absence) of determinism and more on the moral features of the vignettes and questions. Whereas people tend to display compatibilist leanings when asked to make judgments concerning the responsibility of specific agents, when they are asked instead to think about responsibility in the abstract, their intuitions trend toward incompatibilism. There is an ongoing debate about how best to explain these findings (see Feltz et al. 2009; Nahmias 2011; Nahmias & Murray 2011; Sinnott-Armstrong 2008), but Nichols and Knobe take these results as evidence that people have an incompatibilist theory of free will but apply this theory mistakenly when they consider a concrete, emotionally charged, situation.

In addition to the conflicting intuitions identified by Nichols and Knobe, Nahmias, Coates, and Kvaran (2007) found another interesting asymmetry in people's intuitions about free will and responsibility that is especially germane for present purposes. According to this research, people treat explanations of human behavior that are couched in neuroscientific terms differently than they treat explanations couched in folk psychological terms. In order to explore

this issue, Nahmias and colleagues (2007) systematically varied merely the level at which determinism was described and found that it made a significant difference in people's responses. In one study, participants read the following scenario, either in the "neuro case" or the "psych case," which varied only the bracketed words:

> Most respected [neuroscientists / psychologists] are convinced that eventually we will figure out exactly how all of our decisions and actions are entirely caused. For instance, they think that whenever we are trying to decide what to do, the decision we end up making is completely caused by the specific [chemical reactions and neural processes / thoughts, desires, and plans] occurring in our [brains / minds]. The [neuroscientists / psychologists] are also convinced that these [chemical reactions and neural processes / thoughts, desires, and plans] are completely caused by our current situation and the earlier events in our lives, and that these earlier events were also completely caused by even earlier events, eventually going all the way back to events that occurred before we were born.
>
> So, if these [neuroscientists / psychologists] are right, then once specific earlier events have occurred in a person's life, these events will definitely cause specific later events to occur. For instance, once specific [chemical reactions and neural processes / thoughts, desires, and plans] occur in the person's [brain / mind], they will definitely cause the person to make the specific decision he or she makes. (Nahmias et al. 2007, 224)

Although a *minority* of participants said that people would have free will (38%), be morally responsible (41%), or deserve praise and blame if the *neuroscientists* were right, a substantial *majority* said that people would have free will (83%), be responsible (89%), and deserve praise and blame if the *psychologists* were right.

A plausible explanation for these results is that most people see no conflict between determinism and free will or responsibility when the folk psychological framework remains in place, as it does in the "psych case." But when the causes of our decisions are described in the reductionistic and mechanistic language of neuroscience, many people interpret that to mean that our conscious beliefs, desires, and plans are not playing a causal role in our decisions. That is, many people likely interpret neuroscientific descriptions of our decision making in terms of *bypassing*.[7] If our decisions and actions are produced by our psychological makeup, then people think we are responsible for them, even if

7. It is worth pointing out that there are at least two possible kinds of bypassing: (a) weak bypassing, whereby our *conscious* mental states play no etiological role, but maybe our *unconscious* mental states still do play a role; and (b) strong bypassing, whereby *neither* our conscious *nor* our unconscious mental states play an etiological role. The reductionistic and mechanistic

our psychological makeup itself is completely caused by prior events. But if our decisions and actions are produced by chemical and neural processes in our brains, then many people interpret that as an explanation that competes with a folk psychological explanation. It is not determinism per se that is threatening their beliefs in free will and responsibility, but rather there appears to be something unique about the kind of reductionistic and mechanistic explanations one finds in neuroscience that influences people's intuitions about free will and responsibility.

Two other research projects lend further support to these descriptive claims about the impact of neuroscience on these intuitions. First, recent studies by Nahmias and Dylan Murray suggest that when people take determinism to threaten free will and moral responsibility, most do so because they misinterpret determinism to involve bypassing (Nahmias & Murray 2011). In these studies, participants read a variety of different descriptions of determinism, and across these cases, most of those who interpreted these descriptions to threaten free will and responsibility also interpreted them to mean that people's beliefs, desires, and decisions have no effect on what they end up doing. Indeed, people's responses to the questions about bypassing statistically mediated their responses to questions about free will, responsibility, and blameworthiness, providing evidence that bypassing, and not determinism per se, is doing the causal work of mitigating judgments of freedom and responsibility. So, although determinism, properly understood, does not entail that our mental states have no effect on what we do, certain descriptions of determinism or causation seem to prime people to mistakenly assume that it does. One way to prime the bypassing mistake is with neuroscientific or reductionistic explanations that are taken to compete with folk psychological explanations. If neural processes completely explain our actions, then what causal work is left for our beliefs and desires to do? This "competition" between levels of explanation will be especially salient to people who think that the mind and brain are distinct substances or that mental processes cannot be understood in terms of neural processes.

Another research project that supports the descriptive claim that neuroscience may pose problems for the law comes from recent work showing that diminishing people's belief in free will can change their behavior. For instance, when people read a passage by the neuroscientist Francis Crick that tells them "you are nothing but a pack of neurons," or a series of statements, such as "Every action that a person takes is caused by a specific pattern of neural

language of neuroscience sometimes seems to suggest strong and not just weak bypassing. Figuring out whether people take advances in neuroscience to entail strong rather than merely weak bypassing is an issue that calls for more systematic testing.

firings in the brain," then they are more likely than controls to cheat and to lie (Vohs & Schooler 2008), and they are less likely to help others and more likely to be aggressive toward others (Baumeister et al. 2009). After reading such primes, participants are less likely to agree with statements that affirm free will and responsibility. There are various explanations for these effects. But the relevant point for present purposes is that the primes used in these studies present the threat to free will using scientific, reductionistic, often specifically neuroscientific, explanations of human behavior. And these primes have effects on people's responses to statements about free will, on their social behavior, and on whether they hold others responsible. Again, it appears that regardless of whether people *should* interpret neuroscience as a potential threat to free will and responsibility, they *do*.

As we discussed earlier, one area in which these issues are especially germane is mental health law. Researchers are making great strides in uncovering and understanding neurobiological causes of abnormal thought and behavior, and their findings are increasingly making their way into both the popular press and the courtroom. Given that the law already correctly views mental illness as something that sometimes bypasses normal human cognition and behavior, explanations of mental illness that are couched in the language of neuroscience might be especially likely to influence the intuitions and judgments of legal decision makers, including judges and juries, regarding responsibility and punishment. In the following section, we will look at the findings from new studies that probe folk intuitions about the relationships between the mind, the brain, and the law.

2. THE MIND, THE BRAIN, AND THE LAW: SOME NEW STUDIES

As neuroscience continues to advance, it may be possible to use brain imaging to identify, diagnose, and explain some psychiatric illnesses (e.g., Caspi & Moffitt 2006; Farah 2002). This prospect, though not yet fully realized, is likely to be accompanied by a host of challenging legal and psychological issues. One in particular is that neuroscientific explanations of mental illnesses may conflict with the way that people ordinarily think about behavior and its causes. This possibility is particularly pertinent in the legal domain, in which judgments of wrongdoing hinge on everyday mental concepts (folk psychology). Whereas ordinary psychological explanations of wrongdoing tend to focus on the causal role of mental states such as intentions, beliefs, and desires, neuroscientific explanations instead focus attention on the causal role of physiological states within the brain. Although these explanations may be seen as compatible, emphasizing the role of brain states (as opposed to mental states) may nonetheless differentially affect the way individuals judge wrongdoers. In particular,

because brain-based explanations emphasize the causal power of mechanistic, physiological events instead of people's inner mental life (their beliefs, desires, and intentions) that are informative of their true characters (see Pizarro & Tannenbaum 2011; Sripada 2010), brain-based explanations may diminish the perceived culpability of criminal wrongdoers, in turn lessening the punishment that is deemed appropriate for them. It's unclear at this stage whether these brain-based explanations diminish perceived culpability because they are thought to bypass all mental states, because they are thought to bypass character (i.e., long-lasting dispositions to act), or because they are thought to bypass the deep self (even if not all mental states). Some of our recent research has further explored these possibilities, with a particular emphasis on how judgments of culpability and punishment are affected by the inferences people make about the connection between a wrongdoer's bad actions and his character (Gromet et al. 2011).

In these studies, participants evaluated wrongdoers who committed intentional actions that caused another person's death. In each case, the wrongdoer's actions were caused by an emotional dysfunction that was described either as a product of the mind (mental/psychological) or as a product of the brain (neural/neurological). These descriptions were identical except for the framing of the dysfunction in neurological or psychological terms. For example, in one study, following a description of a killing carried out by a man named Gary, participants read that he suffered from an emotional dysfunction that impaired his ability to regulate his anger. We presented the following expert testimony that varied whether mental or neural terms were used to describe the wrongdoer's dysfunction, with the information in square parentheses varied between subjects:

> This [psychological/neurological] dysfunction impairs a person's ability to regulate his anger and control his violent impulses. The expert also testified that as a result of this [psychological/neurological] dysfunction, Gary tends to overreact in social situations. Furthermore, his [psychological/neurological] dysfunction makes it very difficult for Gary to control his behavior in these situations, which can lead to violent outbursts.

Framing an offender's action as caused by a dysfunction, described in either psychological or neurological terms, reduced his perceived culpability compared with control conditions, in which no dysfunction was mentioned. But framing a wrongdoer's dysfunction and behavior in neurological terms had a mitigating effect on people's assessments of his culpability (as well as blameworthiness and responsibility) and their judgments about appropriate punishment for him, compared with framing his dysfunction in psychological terms. We found that

this effect holds regardless of whether the dysfunction is described as due to an environmental cause (i.e., childhood abuse) or a genetic cause (a genetic predisposition). Moreover, this mind-versus-brain difference appears to arise because brain-based framings decouple the wrongdoer's actions from his true character. In the study described above, for instance, when the neurological framing was used, participants rated Gary's actions as being less reflective of whom he is as a person and of his true character than when the psychological framing was used. Essentially, neurological information about a wrongdoer's dysfunction is seen as less diagnostic of the wrongdoer's character than the same information framed in psychological terms. We have examined alternative accounts for this mitigating effect of brain-based explanations, including whether the different framings produce differences in the offender's perceived control over his actions, in the plausibility of the dysfunction causing behavior, or in the credibility of expert testimony. However, none of these alternative accounts have received strong support.

To further corroborate the role of character in underlying the extra mitigating power of brain-based explanations, an additional study showed that the mitigating effect of neuroscientific information was eliminated when there was already a strong connection between a wrongdoer's actions and his moral character. Independent of the mind-versus-brain difference, this study experimentally manipulated how diagnostic the wrongdoer's criminal behavior was of his overall character, by varying whether the wrongdoer had a strong preexisting reason to desire his eventual victim's death. When no such preexisting desire existed, consistent with the previous studies, participants viewed the wrongdoer whose dysfunction was framed as neurological to be less culpable and less worthy of punishment than the wrongdoer who had the identical dysfunction framed in psychological terms. However, when the wrongdoer did have a strong preexisting desire for his victim's eventual death, participants viewed the wrongdoer's actions as highly reflective of his character, regardless of whether his conduct was framed in psychological or neurological terms. And, based on this judgment, participants then no longer viewed brain-based explanations as more mitigating than mind-based explanations.[8]

In sum, these findings illustrate that people make different inferences based on whether criminal behavior is explained in terms of mental states or brain states, and these inferences influence people's judgments of responsibility and punishment. When people are presented with behavior that is described as a product of the brain, this information serves to reduce the connection that people see between how an offender acted and who he is as a person, thus leading to reductions in perceived responsibility and punishment. If future research

8. This conclusion also seems supported by Woolfolk et al. 2006.

yields similar results, we may need to rethink the role that neuroscience will play in determinations of criminal responsibility during legal trials. But before we talk about some of the potential implications of our research on people's views concerning neuroscience and responsibility, we will first offer what we take to be the best explanation of the gathering data.

3. NEUROSCIENCE AND THE DEEP SELF

In social psychology, one of the most influential approaches to understanding judgments of moral responsibility derives from the work of Franz Heider, in his 1958 classic, *The Psychology of Interpersonal Relations*. Heider conceived of judgments of moral responsibility as consisting of an ordered sequence of progressive stages. Early stages draw a *physical* link between the agent and outcome (Did the agent cause the outcome?), whereas later stages draw a *volitional* link (Did the agent foresee that the outcome would occur? Did the agent desire the outcome? Did the agent intend the outcome?). Progression through each stage marks a "tighter" link between the agent and the outcome, and the degree of assessed moral responsibility commensurately rises. Subsequent theorists in social psychology have elaborated on Heider's stage theory in various ways but have largely kept the overall structure (see Shaver 1985; Schlenker et al. 1994). For our purposes, the most import feature of these Heiderian-inspired models is the central role accorded to the latter stages comprising the volitional link, and in particular the role of action-directed mental states such foresight, desire, and intent. According to what we can call *volitionist models* of moral responsibility, assuming the appropriate causal link between the agent and outcome has been established, then if the agent *desires* the outcome, *intends* to bring about the outcome, and *foresees* that his doing the action will cause the outcome to occur, then the agent is morally responsible for the outcome.

In philosophy, models broadly similar to the kinds of volitionist models encountered in social psychology have also been proposed (see Levy & McKenna 2009 for a discussion). However, in addition, there is another family of accounts of moral responsibility that is highly influential in philosophy, but has not been developed much in psychology. These "deep self" accounts of moral responsibility provide a unique vantage point for understanding how neuroscience affects moral responsibility judgments, so we will now sketch the motivation and structure of deep self models.

There are many different kinds of deep self accounts of moral responsibility in the philosophical literature (see Frankfurt 1971; Smith 2008; Watson 1975; see Wolf 1990 for a seminal discussion). What these accounts all have in common is that they draw a basic distinction within the set of an agent's conative attitudes, that is, the set of her desires, wants, values, and other motivationally

relevant states. More specifically, deep self models distinguish between conative attitudes that are "surface" and those that are "deep." When a person performs an action, there are invariably certain *surface attitudes* that help to explain why the action was performed—states such as desires and intentions. These mental states are quite *specific*, that is, they are directed at a particular action at a particular time, and *temporary*, that is, they typically arise before the action and dissipate once the action is performed. In addition to surface attitudes such as desires and intentions, however, people also have *deep attitudes* that are more fundamental and important. This is reflected in the fact that we ordinarily use terms such as "values," "cares," and "core commitments" to refer to these kinds of attitudes. Deep self theorists disagree about what makes these states more central and significant, with one theory positing that these states bear a distinctive connection to practical reasoning (Watson 1975), while another theory emphasizes a role for higher order attitudes that endorse one's first-order motives (Frankfurt 1971). But setting these differences aside, these theorists agree that a person's deep attitudes are the very essence of who she is as a practical agent, the *self* that underlies all her actions, and thus these attitudes should play a central role in how she is assessed for what she does. Theories of moral responsibility in psychology, such as the volitionist theories discussed earlier, tend to emphasize the role of surface attitudes in people's judgments of moral responsibility. On this view, a person is morally responsible for an action if she has the appropriate desires and intentions. Deep self theories, in contrast, claim that in addition to surface attitudes, deep attitudes also make important contributions to moral responsibility judgments, in ways to be detailed below.

In a mentally healthy person who is not under duress or constraint, deep attitudes, surface attitudes, and actions will usually be in harmony. That is, people tend to perform actions based on (surface) desires and intentions that are for the most part in agreement with their own underlying (deep) values and core commitments. But there are a number of factors—such as ignorance, coercion, constraints, irresistible impulses, addiction, and the like—that can cause an agent's deep attitudes to diverge from her surface attitudes and her actions. In these cases, our judgments about the agent's moral responsibility appear to be highly sensitive to the content of the agent's deep attitudes, thus providing critical evidence for the deep self view.

Harry Frankfurt's example of willing and unwilling addicts illustrates this point (Frankfurt 1971). Consider two addicts, both of whom have an irresistible desire to use a narcotic. The "unwilling addict" *rejects* his addiction, and desires that his desire to use the narcotic be extinguished. The "willing addict" *endorses* his addiction, and were his desire to use the narcotic ever extinguished, he would seek to reinstate it. When each addict uses the drug, the surface attitudes that drive his action, e.g., the irresistible desire to use the narcotic, are the

same. But there is a strong intuition that the addicts differ in terms of moral responsibility; the willing addict *is* morally responsible for his action while the unwilling addict is *not*, or at least the two addicts differ in their degree of moral responsibility. Volitionist models have difficulty making sense of this difference because the two addicts do not differ in terms of the action-directed psychological states (e.g., desire, intention, foresight) that these models claim are determinative of moral responsibility. Deep self accounts of moral responsibility, in contrast, readily explain the difference in responsibility between the willing and unwilling addict in terms of differences in their respective deep selves. The unwilling addict rejects and thus "stands against" his narcotic-directed desires so he is not morally responsible for the resulting actions. The willing addict endorses and thus "stands with" his narcotic-directed desires so he is morally responsible for the resulting actions.

Using *philosophical* theories of the deep self as a starting point, Sripada (2010; 2011; forthcoming) formulated a *psychological* model of intentionality and responsibility judgments. The key element of the model is a *concordance criterion* that specifies how people use information about deep attitudes on the one hand and information about actions and outcomes on the other hand to arrive at judgments about whether the agent is morally responsible for the action or outcome:

> *Concordance criterion for moral responsibility judgments*: An agent is judged to be morally responsible for an action or outcome to the extent that the action or outcome is judged to be concordant with the agent's deep self.

Sripada's "deep self concordance account" raises a number of questions. For example, how do people decide which of an agent's attitudes are truly deep and which are not? Also, how should the notion of concordance be understood? In particular, it seems we can distinguish *wide* versus *narrow* notions of concordance. Wide concordance requires that the action in question promotes the agent's *overall set* of values, cares, and core commitments. Narrow concordance, in contrast, only requires that the action in question promote *some or other element within* the agent's set of values, cares, and core commitments. Preliminary data suggest that it is specifically the narrow notion of concordance that is relevant to people's moral responsibility judgments (Sripada unpublished). Finally, the Deep Self Concordance Model raises intriguing questions about what happens in cases of "deep conflict" when an agent appears to simultaneously hold diverging deep attitudes. A full discussion of these issues is beyond the scope of this chapter. Indeed, most of the preceding issues are only just beginning to be addressed by philosophers and psychologists, and much further research is required.

Although deep self theory is relatively new to the psychological scene, and many questions about how to understand the approach remain to be resolved, it provides a promising framework for addressing questions about how neuroscience will affect people's view of moral responsibility. The current literature tends to see the presumed conflict between neuroscience and moral responsibility in abstract, metaphysical terms—people perceive that neuroscience undermines moral responsibility because it shows that people's choices are produced by neural-cum-physical mechanisms that are subject to the, presumably deterministic, laws of physics (see Greene & Cohen 2004; Nichols & Knobe 2007). Deep self theory, in contrast, portrays the impact of neuroscience along different, distinctively *nonmetaphysical*, lines. On the deep self view, the tension between neuroscience and moral responsibility arises because the kinds of mechanisms that neuroscience identifies as being the source of our actions, at least in some cases, turn out to be of the wrong sort to ensure that our deep attitudes, that is, our underlying values, cares, and commitments, are appropriately reflected in our actions.

To illustrate this idea, consider a recent experiment by Nadelhoffer and colleagues (in preparation). All subjects in the experiment were told of the case of John Smith, who "got into an argument with a co-worker" and subsequently "pushed the [co-worker] to the ground and struck the [co-worker] despite being physically unprovoked." One third of the subjects were in the "neuroscience" condition and read testimony from a neuroscientist who stated that, based on a detailed examination of *brain-based* information, Smith is in the highest risk category for violently reoffending within 5 years. The remaining subjects were either in the "actuarial" or "psychological" condition, where they read testimony from a statistician and a psychologist, respectively, who, based on their respective non-neuroscientific sources of information, also reported that Smith is in the highest risk category for violently reoffending within 5 years.

Results of the study showed that subjects assigned significantly more punishment to Smith in the neuroscience condition compared with the actuarial and psychological conditions. Moreover, attributions of attitudes to Smith's deep self mirrored the pattern of responsibility, blame, and punishment judgments. That is, people judged that Smith's action was more reflective of his true underlying values and character in the neuroscience condition compared with the other two conditions. The precise reasons that neuroscientific descriptions produced altered perceptions of the deep self are unclear. We speculate that accounts of deviant behavior, when pitched in neuroscientific terms, may be viewed as more credible, or the person may be viewed as more fundamentally and/or permanently affected. These perceptions would lead lay observers to perceive that the person's actions reflect the kind of person whom he truly is. Additionally, people may assume that brain-based disorders have more

pervasive and comprehensive effects, so that it is not just the person's outward behavior but also his inner values and core commitments that are disrupted. Further research is needed to test these speculations.

The preceding experiment illustrates that the impact of neuroscience on moral responsibility judgments need not always be mediated by abstract metaphysical doctrines such as determinism or mechanism. Rather, neuroscience's impact on responsibility judgments might, at least in some cases, be mediated by construals about whether the agent's deep self is or is not the source of the person's behavior. In addition, it is noteworthy that in this experiment, there is no obvious reason for subjects to have supposed that Smith's *surface mental states*, that is, his means-end beliefs, surface desires, and intentions, differed across the three conditions of the experiment. Thus, this experiment highlights that neuroscientific explanations might have a relatively *specific* impact on construals of the deep self. In contrast, surface mental states such as desires, means-end beliefs, and intentions—the kinds of action-directed mental states that volitionist theories say are relevant for assessing moral responsibility—may be largely unaffected.

Results from the preceding study extend and nuance the results from the studies we reported earlier (section 2) on how neuroscientific information affects judgments of moral responsibility. These studies also illustrate how deep self theory provides an alternative, complementary framework for studying the impact of advances in neuroscientific knowledge on folk practices of praise and blame. Psychiatric neuroscience is increasingly uncovering the brain-based mechanisms that produce abnormal patterns of behavior in disorders such as depression, mania, schizophrenia, and addiction. Deep self theory predicts that at least some of the impact of this new knowledge on judgments of moral responsibility and blame will be mediated by people's perception of how the disorder affects the status of the person's deep self. In some cases, people will perceive that these disorders sever the usual connection between a person's deep self and her actions, causing her to produce actions that she, *deep down inside*, does not endorse, thus mitigating blame. In other cases, people might perceive that these disorders are associated with a more thoroughly compromised deep self, in which case blame will be enhanced. It will require coordinated research by philosophers and psychologists to test these predictions, to determine what makes people go one way in some cases and the opposite way in other cases, and to evaluate their import for the law and related social institutions.

CONCLUSION

In this paper, we aimed to shed new light on the growing debate among philosophers, psychologists, and legal scholars concerning the potential influence

that advances in neuroscience are likely to have on our ordinary understanding of free will and responsibility. We also discussed the results of some new studies that suggest that regardless of whether one believes that neuroscientific discoveries ought to alter our views concerning moral agency, it appears that, descriptively speaking, people view wrongdoers as less morally responsible for their actions when their behavior is described as a product of the brain.[9] That being said, this research only scratches the surface, leaving many questions for further research. But even the research we have carried out so far raises interesting practical and philosophical questions. The fact that people are sensitive to whether and to what extent a person's actions are rooted in his underlying character and values suggests that the law's criteria for determining culpability, which do not include character as a salient factor (e.g., Fletcher 1998), may lead to conflicting intuitions in the minds of jurors, at least when jurors are not given careful instruction.

Our recent studies also illustrate how neuroscience is raising new questions about how we are to conceptualize human agency and about how we judge moral responsibility. Whereas some have argued that neuroscience does not currently pose a threat to notions of moral and legal culpability (see Berker 2009; Morse 2008), others have contended that the whole notion of moral responsibility is challenged by emerging developments in cognitive neuroscience (Greene & Cohen 2004; see also Eagleman 2011). The current studies discussed here do not propose any solution to these difficult normative issues. Whether cognitive neuroscience is poised to usher in a revolution in our understanding of moral responsibility and free will remains to be seen, and that itself is largely an empirical question. In the meantime, we believe that psychologists and philosophers must continue to work collectively in order to better understand the complex relationships between the mind, the brain, and the law.

9. It's worth pointing out that future research needs to explore the possibility that the threat of mechanism as it relates to bypassing is distinct from the threat of mechanism as it relates to the deep self view. After all, if what undermines responsibility is bypassing, then people may be held responsible for shallow desires even if they have no connection to the deep self, at least when the agent has control. In contrast, if what undermines responsibility is lack of connection to the deep self, then people will not be held responsible for shallow desires that have no connection to the deep self. If this is correct, then the bypassing view and the deep self view may actually be *competing* accounts of the gathering data. Although shedding light on this issue is a task for another day, we nevertheless thought it merited mentioning because we have assumed for the purposes of this chapter that the bypassing view and the deep self view are complementary rather than competing.

REFERENCES

Ayer, A. J. 2003. Freedom and necessity. In G. Watson (Ed.), *Free Will.* Oxford: Oxford University Press: 15–23.

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy and Public Affairs* 37: 293–329.

Caspi, A., & Moffitt, T. E. (2006). Gene-environment interactions in psychiatry: Joining forces with neuroscience. *Nature Reviews Neuroscience* 7: 583–590.

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin* 35: 260–268.

Chisholm, R. 2003. Human freedom and the self. In G. Watson (Ed.), *Free Will.* Oxford: Oxford University Press: 26–37.

Clarke, R. 2003. *Libertarian Accounts of Free Will.* Oxford: Oxford University Press.

De Brigard, F., Mandelbaum, E., & Ripley, D. 2009. Responsibility and the brain sciences. *Ethical Theory and Moral Practice* 12(5): 511–524.

Dennett, D. 1984. I could not have done otherwise: So what? *Journal of Philosophy* 81: 553–565.

Double, R. 1991. *The Non-Reality of Free Will.* New York: Oxford University Press.

Eagleman, D. (2011, July/August). The brain on trial. *Atlantic Magazine.*

Ekstrom, L. 2000. *Free Will: A Philosophical Study.* Boulder, CO: Westview Press.

Farah, M. J. (2002). Emerging ethical issues in neuroscience. *Nature Neuroscience* 5: 1123–1129.

Feltz, A., Cokely, E., & Nadelhoffer, T. 2009. Natural compatibilism vs. natural incompatibilism: Back to the drawing board. *Mind & Language* 24: 1–23.

Fischer, J. M. 1994. *The Metaphysics of Free Will.* Oxford: Wiley-Blackwell.

Fischer, J. M., Kane, R., Pereboom, D., & Vargas, M. 2007. *Four Views on Free Will.* Oxford: Wiley-Blackwell.

Fischer, J. M., & Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility.* Cambridge, UK: Cambridge University Press.

Fletcher, G. P. (1998). *Basic Concepts of Criminal Law.* New York: Oxford University Press.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.

Green, T. A. 1995. *Freedom and Criminal Responsibility in the Age of Pound: An Essay on Criminal Justice,* 93 Mich. L. Rev. 93(2): 1915–2053

Greene, J. D., & Cohen J. D. 2004. For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London B* (Special Issue on Law and the Brain) 359: 1775–17785.

Gromet, D. M., Goodwin, G. P., Tang, S., Nadelhoffer, T., & Sinnott-Armstrong, W. P. (2011). *Mind, brain, and character: How neuroscience affects people's views of wrongdoers.* Manuscript in preparation. University of Pennsylvania.

Honderich, T. 1988. *A Theory of Determinism.* Oxford: Oxford University Press.

Kane, R. 1996. *The Significance of Free Will.* New York: Oxford University Press.

Kane, R. 2011. *The Oxford Handbook of Free Will: Second Edition.* Oxford: Oxford University Press.

Levy, N., & McKenna, M. 2009. Recent work on free will and moral responsibility. *Philosophy Compass* 4: 96–133.

Lycan, W. 2003. Free will and the burden of proof. In A. O'Hear (Ed.), *Minds and Persons: Royal Institute of Philosophy Supplement*. Cambridge, UK: Cambridge University Press: 107–122.

Mele, A. 1996. *Autonomous Agents*. Oxford: Oxford University Press.

Morse, S. J. (2008). Determinism and the death of folk psychology: Two challenges to responsibility from neuroscience. *Minnesota Journal of Law, Science and Technology* 9: 1–35.

Nadelhoffer, T., & Nahmias, E. 2007. The past and future of experimental philosophy. *Philosophical Explorations* 10(2): 123–149.

Nadelhoffer, T., Sripada, C, Gromet, D., & Sinnott-Armstrong, W. (in preparation) Neuroscience, Future Dangerousness, and the Criminal Law.

Nahmias, E. 2011. Intuitions about free will, determinism, and bypassing. In R. Kane (Ed.), *The Oxford Handbook on Free Will*. 2nd ed. Oxford: Oxford University Press: 555–576.

Nahmias, E., Coates, J., & Kvaran, T. 2007. Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy* 31: 214–242.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. 2006. Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73(1): 28–53.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. 2005. Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18(5): 561–584.

Nahmias, E., & Murray, D. 2011. Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff, & K. Frankish (Eds.), *New Waves in Philosophy of Action*. New York: Palgrave-Macmillan: 189–216.

Nichols, S., & Knobe, J. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous* 41: 663–685.

O'Connor, T. 2000. *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.

Pereboom, D. 2001. *Living Without Free Will*. Cambridge, UK: Cambridge University Press.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *Herzliya Symposia on Personality and Social Psychology. The Social Psychology of Morality: Exploring the Causes of Good and evil*. Arlington, VA: APA Press.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. 1994. The triangle model of responsibility. *Psychological Review* 101(4): 632–652.

Shaver, K. G. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. New York: Springer-Verlag.

Sinnott-Armstrong, W. 2008. Abstract + Concrete = Paradox. In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy*. New York: Oxford University Press: 209–230.

Smilansky, S. 2000. *Free Will and Illusion*. Oxford: Oxford University Press.

Smith, A. 2008. Control, responsibility, and moral assessment. *Philosophical Studies* 138: 367–392.

Sripada, C. S. Forthcoming. What makes a manipulated agent unfree? *Philosophy and Phenomenological Research.*

Sripada, C. S. 2010. The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies* 151: 159–176.

Sripada, C. S. Forthcoming. The deep self and psychology of excuse.

Stace, W. (Ed.). 1960. *Religion and the Modern Mind.* Philadelphia: Lippincott.

Strawson, G. 1986. *Freedom and Belief.* Oxford: Clarendon Press.

Strawson, P. F. 2003. Freedom and Resentment. In G. Watson (Ed.), *Free Will.* Oxford: Oxford University Press: 72–93.

Van Inwagen, P. 1983. *An Essay on Free Will.* Oxford: Oxford University Press.

Vohs, K., & Schooler, J. 2008. The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science* 19: 49–54.

Watson, G. 1975. Free agency. *Journal of Philosophy* 72: 205–220.

Watson, G. 2003. *Free Will.* 2nd ed. Oxford: Oxford University Press.

Wolf, S. 1990. *Freedom Within Reason.* Oxford: Oxford University Press.

Woolfolk, R. L, Doris, J. M., and Darley, J. M. 2006. Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition* 100: 283–301.

OXFORD SERIES IN NEUROSCIENCE, LAW, AND PHILOSOPHY

SERIES EDITORS

Lynn Nadel, Frederick Schauer, and Walter P. Sinnott-Armstrong

Conscious Will and Responsibility
Edited by Walter P. Sinnott-Armstrong and Lynn Nadel

Memory and Law
Edited by Lynn Nadel and Walter P. Sinnott-Armstrong

Neuroscience and Legal Responsibility
Edited by Nicole A. Vincent

A Primer on Criminal Law and Neuroscience
Edited by Stephen J. Morse and Adina L. Roskies

The Future of Punishment
Edited by Thomas A. Nadelhoffer

The Future of Punishment

EDITED BY THOMAS A. NADELHOFFER

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging-in-Publication Data
The future of punishment / edited by
Thomas A. Nadelhoffer.
pages cm.—(Oxford Series in Neuroscience, Law, and Philosophy)
ISBN 978-0-19-977920-8 (hardback : alk. paper)—
ISBN 978-0-19-977935-2 (e-book) 1. Criminal justice, Administration of—Moral and
ethical aspects. 2. Punishment—Moral and ethical aspects. 3. Responsibility.
4. Free will and determinism. I. Nadelhoffer, Thomas, editor of compilation.
HV7419.F885 2013
364.601—dc23
2012029161

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper