2013

Skepticism Concerning Human Agencies: Sciences of the Self Versus 'Voluntariness' in the Law

Paul Sheldon Davies
College of William and Mary, psdavi@wm.edu

Recommended Citation

Vincent, N. (2013). Neuroscience and legal responsibility (Oxford series in neuroscience, law, and philosophy). New York: Oxford University Press.

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

Neuroscience and Legal Responsibility

EDITED BY NICOLE A VINCENT

OXFORD
UNIVERSITY PRESS

5

Skepticism Concerning Human Agency

Sciences of the Self Versus "Voluntariness" in the Law

PAUL SHELDON DAVIES

The findings of neuroscience cast grave doubts on the view of human agency implicit in the law. They do this by forcing us toward a form of skepticism concerning our capacities as agents. That is the thesis of this chapter.

The findings of neuroscience do not cast doubt in isolation. They do so when combined with findings in cognitive and social psychology and findings in evolutionary theory and primate cognition, and when integrated into a large-canvas view that sometimes results from informed philosophical reflection. This is a powerful methodological directive demonstrated throughout *On the Origin of Species*, a directive that ought to be adopted in the study of the self as much as in the study of life.

The logic of Darwin's (1859) argument is a sequence of abductive arguments concerning a broad range of distinct biological, geological, and geographical phenomena. None of his arguments is decisive taken alone, and some are stronger than others, but their combined power comes from a weighty convergence upon a single, unifying view of life, drawn from an accumulation of inferences to the best explanation concerning several distinct phenomena. This strategy, so potent in the study of life, is our best bet in the study of capacities that animate living things. Questions about human agency, for instance, including the viability of concepts of legal responsibility, cannot be settled

with a small set of experiments or a localized hypothesis. What we need is a large-canvas view that integrates knowledge from the relevant sciences. And once we formulate such a view, we find that at present we do not know what kind of agent we are. An informed skepticism best describes where we are today.

The shape of my argument is as follows. Section I introduces the main target of my discussion: a concept of voluntariness that appears essential to a concept of criminal responsibility. I focus on "voluntariness" for the sake of concreteness. Once we appreciate the converging doubts against this concept, doubts concerning other concepts of agency naturally arise. Section II is a brief summary of my very general grounds for thinking that the methods with which we study the human self are in need of reform. The following two sections offer a more specific defense of this call for reform, as well as a few preliminary reformative steps, what I call *directives for inquiry*. I propose one directive in section III that is ameliorative or curative in nature and three additional directives in section IV that are exploratory rather than curative. Then, in section V, on the basis of my proposed directives, I defend my skepticism regarding human agency. I conclude by drawing out the implications of this skepticism for the specified notion of criminal responsibility.

"VOLUNTARINESS" AND CRIMINAL RESPONSIBILITY

For the sake of concreteness, I focus on the partial characterization of criminal guilt from section 2.01 of the Model Penal Code, which states that an agent is criminally guilty for a given action only if it was voluntary, where "voluntary action" is characterized in conditions (a) to (d):

> A person is not guilty of an offense unless his liability is based on conduct that includes a voluntary act or the omission to perform an act of which he is physically capable. The following are not voluntary acts within the meaning of this Section: (a) a reflex or convulsion; (b) a bodily movement during unconsciousness or sleep; (c) conduct during hypnosis or resulting from hypnotic suggestion; (d) a bodily movement that otherwise is not a product of the effort or determination of the actor, either conscious or habitual.

The core assertion is simple, at least on the surface: the attribution of guilt for an action is justified only if the agent's conduct was a bodily movement produced by "the effort or determination of the actor, either conscious or habitual."[1]

It may appear, however, that a person can be guilty in a quite different way, by failing to perform some action despite being physically capable. An agent may be guilty not by virtue of actions that result from effort or determination but simply by virtue of omissions, in which case this section of the Model Penal Code may be interpreted as articulating two distinct concepts of legal responsibility, only one of which employs "voluntariness." Although I am skeptical of any such "two concepts" interpretation, I shall, for the sake of this discussion, restrict my argument to acts of commission and silently pass over the question of whether criminal acts of omission rest upon a prior "voluntary" act that the agent performed or reasonably should have performed.[2] After all, if the single notion of legal responsibility applied to acts of commission falls to my skepticism, that is enough to show that the concept "voluntariness" in the Code is deeply problematic.

Note, then, that the characterization of "voluntariness" in (a) to (d) is remarkably uninformative. The first three conditions are entirely negative: bodily movements not "determined" by the agent—reflexive movements, sleepwalking, for example—are not voluntary. What, then, are the distinguishing properties of movements that are voluntary? We are not told. Condition (d) merely generalizes from the negative characterization in (a) to (c): only movements produced by the effort or determination of the actor are voluntary.

That this section of the Code is nonspecific is not an automatic indictment, however. Laws are tools designed to fulfill certain functions, and some functions can be executed with relatively blunt instruments. This, I surmise, is true of the above characterization of voluntariness. The relevant conditions are sparsely specified on the assumption that there is enough shared cultural knowledge concerning the causes of human conduct to fill the gaps. The lack of specificity in the law is tolerable, perhaps preferable, because our shared cultural knowledge enables us—lawyers, judges, and jurors—to apply the law in light of the particulars of each case. This may provide a degree of flexibility that a fuller specification of "voluntariness" may rule out.

If the above characterization of voluntary action is deliberately generic in this way, then the crucial assumption must be something like this: most adult citizens (including those likely to serve as jurors) know that we are agents who sometimes "determine" their actions and also know when, under what conditions, our actions are in fact the results of our "determinations." If this crucial assumption is false, then the law cannot fulfill its function. If there is not shared knowledge in areas where the law has jurisdiction—if lawyers, judges, and jurors do not know enough to reliably discern actions genuinely determined by the actor from those determined by other factors—then the law is defective.

The question, then, is whether this crucial assumption is true. Do most adult citizens know that we are agents who sometimes determine their actions? Do most know when, under what conditions, our actions result from such determinations? The question is not whether most citizens believe that they have such knowledge, but whether they in fact have it. To answer this question, we have no recourse but to turn to our best developed scientific theories of the self and, on the model of Darwin in the *Origin*, paint in vivid colors our most informed large-canvas view of our capacities as agents. Once we do that, we will see that the answer to this question, in light of current knowledge, is a decidedly negative one; the concept of voluntariness in the above notion of legal responsibility is at odds with what we know about ourselves.

THE AIMS AND STRATEGIES OF CONTEMPORARY THEORIES OF THE SELF

I turn to the general aims and strategies of contemporary theories of the self. I do not claim that these aims and strategies are explicitly endorsed by contemporary theorists in philosophy, psychology, or legal studies. I claim only that they accurately reflect the overarching commitments and methods of many theorists in those areas. As we will see, these aims and strategies tend to diminish rather than enhance our chances of discovering the truth about ourselves. The methods with which we study ourselves are in need of reform.[3]

At a high level of abstraction, the methods with which we study ourselves are either *conceptually conservative* or *conceptually imperialistic*. Neither conservatism nor imperialism by itself is objectionable but, when applied to *dubious concepts*, both methods retard our efforts at discovering the truth. The call for reform, in consequence, is a call for directives that cure us of conservatism and imperialism, as well as directives that guide us in our efforts to discover the truth.

Conceptual conservatism is a strategic orientation toward inquiry, affecting the way we frame our questions and answers. The overarching goal is to conserve or save as far as possible concepts of apparent importance, concepts that appear salient in our more general worldview. The preferred strategy for saving apparently important concepts is to "locate" them amid the concepts and claims of some preferred base theory. For naturalists, the preferred base is usually a well-developed scientific theory; for non-naturalists, it is some well-entrenched part of our inherited worldview. Ruse (2003) presents a book-length exercise in conceptual conservatism, aspiring to save a concept of normative functions in evolutionary biology even at the cost of resuscitating Kant's (1790) theory of natural purposes.

Conceptual imperialism is more ambitious than conservatism. The overarching goal is not to save apparently important concepts as far as possible but to force the rest of our conceptual scheme to accommodate certain concepts at any cost. Certain concepts, it is assumed, have dominion over other concepts and over methods of inquiry. These, according to the imperialist, are concepts without which we would be unable coherently to think or articulate a view of the relevant phenomena. Chisholm (1964) was an imperialist regarding the human self—he insisted that the libertarian concept of free will had to be retained even at the cost of accepting that every conscious, rational person is a little Thomistic god.

Conservatism and imperialism may be appropriate in some contexts but not when applied to dubious concepts. A concept is dubious when there are justified grounds for excluding it from our theorizing. Such concepts fall into two general groups. Some are *dubious by descent.* These are categories that descend to us from a worldview we no longer regard as true or promising, that have not been vindicated in any well-confirmed theories, but that nonetheless tend to influence the way we frame our inquiries. The concept of a nonphysical soul is illustrative. So far as we can surmise, the neural processes implementing human thought and action operate under the principle of causal closure. Our best evidence for this is the utter lack of experiments in which the best explanation of observed phenomena requires the postulation of a nonphysical cause.

Some concepts are *dubious by psychological role.* These are categories controlled by conceptualizing capacities prone to abundant false positives or false negatives. Consider by analogy visual illusions. Under a range of conditions, our visual capacities produce systemic errors. Similarly, under a range of conditions, our cognitive and affective capacities produce systemic errors. But there is a crucial disanalogy. Most visual illusions are easily identified and compensated for, whereas most conceptualizing illusions occur without the agent's notice. Some conceptualizing illusions are so much a part of our deliberative field that it never occurs to us to be troubled by them—not, at any rate, until we meet with an ingenious experiment that reveals the systemic error.

The much-discussed theory of Daniel Wegner (2002) is a case in point. The human mind, according to Wegner, comprises a system that generates the felt experience of consciously willing and thereby consciously controlling our actions. Not all of our actions, of course, because many of our actions are relatively thoughtless—just those we regard as the products of our will. What is provocative, and what reveals a tendency towards systemic error, is evidence adduced by Wegner that this system of conscious willing operates independently of the low level, nonconscious mechanisms that actually cause us to act.

The mechanisms that cause our actions, it appears, are not the mechanisms that give us certain beliefs and feelings about the causes of our actions. We are led astray by the very constitution of our psychology.

Wegner's theory, when integrated with theories from distinct areas of inquiry, wields considerable power. Indeed, Wegner combines experiments from his lab with evidence (some explicated later) concerning a broad range of affective and cognitive phenomena. The strength of his theory rests upon the integrated view of the self that emerges from these diverse phenomena, especially the view of our apparent capacity to control our actions by consciously willing. It is the breadth of converging evidence that makes it rational to hold that the concept of conscious willing, because it generates an abundance of false positives and negatives that are difficult to detect, is dubious by psychological role.

I will explicate Wegner's theory in due course, but I wish to highlight a general feature of our inclination toward conservatism and imperialism, namely, that we tend to be most conservative or imperialistic with respect to concepts most dubious. The greater the staying power of a conceptual category, the greater our tendency to try to save it, perhaps because we feel confident that long-lived concepts must be tracking something real and important. Indeed, the mechanisms that give concepts their staying power are among the very mechanisms that render some concepts so dubious. Some are preserved by culturally instituted mechanisms of transmission; some by the architecture of our psychology (mechanisms that produce persistent errors we tend not to notice); and some, no doubt, are preserved both by cultural and psychological factors.[4] Concepts preserved by any of these mechanisms are going to recur in our deliberative activities; they are going to appear important precisely because they are so tenacious.

Thus, we must be especially cautious in the study of the human self because the concepts with which we understand our capacities as agents are among the most dubious. Concepts such as "free will" and "moral responsibility" are clearly dubious by descent, thanks to our largely theological ancestry, and many of the concepts with which we understand our capacities as agents are dubious by psychological role, as we are about to see.

CURATIVE DIRECTIVES

If the above line of reasoning is correct, if the need for reform in our methods is real, we must diminish the retarding effects of our conservatism and imperialism regarding dubious concepts. To that end, I propose we adopt directives for inquiry formulated in light of our best theories of the very mechanisms that lead us astray. I begin with a directive designed to diminish the ill effects of concepts dubious by psychological role:

DP: For any concept dubious by psychological role, do not make it a condition of adequacy on our theories that we "save" or otherwise preserve that concept; rather, identify the conditions (if any) under which the concept is correctly applied and withhold antecedent authority from that concept under all other conditions.

Withholding antecedent authority from a concept comes to this: we frame our inquiries without that concept. This is not to adopt eliminativism concerning dubious concepts. The directive is to withhold dubious concepts from inquiry until we have reasonable knowledge of the conditions, if any, under which they can correctly be applied. A further aim is to cultivate intellectual creativity. The aim is not merely to avoid concepts that demonstrably lead us astray but also to put ourselves under pressure to create alternative categories with which to explain and predict the phenomena.[5]

When our knowledge of the mechanisms involved in the application of a concept gives rise to such doubts concerning that concept, the directive in DP is essential. We need a reliable process with which one part of our psychology can mitigate the ill effects of another part. To illustrate, consider a few details of our apparent capacity for consciously willing our actions. Wegner's proposed system is triggered when we consciously perceive instances of the following pair:

A thought about or the intention to perform action A
&
THE perception or recollection of oneself performing action A

You think about taking another sip of wine and then perceive yourself sipping. These conscious inputs trigger an interpretive system in your psychology, the function of which is to render your actions intelligible. It achieves this in two steps. It first produces a causal hypothesis to the effect that you, by virtue of your prior thought or intention, caused yourself to perform the action. The hypothesis is that your conscious thought is the means by which you controlled the production of your action. Then the system produces an accompanying affect, a felt sense of achievement, what Wegner calls the emotion of authorship.[6] This interpretive system does all this even though your action was caused by a separate set of mechanisms. This is Wegner's theory of *apparent mental causation.*

The power of this theory derives from the breadth of additional theories with which it integrates. Consider, for instance, the *forward model* of motor control.[7] Suppose I ask you to perform a simple intentional action. I ask you to touch the tip of your nose with your right pointer finger. As you move your right arm, your brain generates a continuous stream of predictions about

where your arm ought to be at the next instant, relative to the goal of reaching the tip of your nose. That is why it is a "forward" model: the system generates predictions concerning the ideal future location of your arm. These predictions are useful because they are compared to the continuous proprioceptive feedback regarding the actual position and trajectory of your arm. Any mismatch between predicted and actual position is then used to update the signals sent to your muscles, thereby correcting your action in real time. All of this happens with breathtaking speed at a level of processing inaccessible to conscious awareness.

What is intriguing is that, in the course of executing these anticipatory functions, your brain suppresses its own ability to fully process incoming sensory information. Put generally, the processing of sensory information is suppressed or at least attenuated whenever we act intentionally. This appears clear from experiments reported in Blakemore, Wolpert, and Frith (1999).[8] In one study, experimenters first asked subjects to touch the palm of their right hand using a device manipulated with their left hand; subjects were asked, that is, to perform a simple intentional action. They then asked subjects to allow them, the experimenters, to touch subjects' right hand by manipulating the intervening device. The results were striking. When the experimenter initiated the action—when the act of touching was not intended by the subjects—subjects rated the sensation in their right palm as intense and tickly. When, however, subjects initiated the action themselves, they rated the sensation as less intense and tickly. Our motor control capacities, in the course of executing intended actions, suppress the processing of incoming sensory information.

The same attenuation occurred in a second study. Subjects once again were asked to touch the palms of their right hands by manipulating an intervening device with their left hands. What subjects did not know was that the experimenters were introducing very short delays in the operation of the device. With each trial the motion of the intervening device and the subsequent sensory perception were delayed relative to the subjects' initiation of the intended action. The results were striking. When the delay was short, subjects reported that the sensation was neither intense nor tickly. The brain, while executing the intended action, suppressed the processing of sensory input. But as the delay grew longer, as sensation became increasingly distant from action initiation, the sensation of being tickled increased. The sensation was increasingly processed as coming from something external to the self.[9]

This attenuation of sensory information is important in two ways. First, our motor capacities suppress an enormous quantity of sensory information whenever we act intentionally. One part of our psychology (motor control) conceals from another part (conscious awareness) a large set of causal information. And this contributes to what we might call a kind of *phenomenological*

quiet, a degree of subjective silence against which the things that do come to conscious awareness—including the conscious inputs that trigger Wegner's interpretive system—appear salient in our conscious, deliberative fields. The factors that come to conscious awareness, against this backdrop of quiet, are bound to strike us as causally efficacious, especially when processed by an interpretive system dedicated to causal intelligibility. Second, the misleading effects produced by this phenomenological quiet arise from the very architecture of our psychology. This is no small point. Some of Wegner's critics try to dismiss his view by insisting that he is concerned with oddball illusions or marginal mistakes that our otherwise veridical capacities do not suffer. But integrating Wegner's view in this way shows that these critics are mistaken. We are seduced not at the margins but by capacities at our agential core.

Now consider the theory of *naïve realism*, which also integrates with the theory of apparent mental causation. The main elements of naïve realism are three: (1) We tend to assume that we see things in an unmediated and objective manner. (2) We tend to assume that other rational persons will see things as we do. (3) We tend to dismiss those who disagree as ignorant, slothful, irrational, or biased. The background suggestion is that, because each of us approaches a situation, especially situations involving other persons, with limited knowledge and extensive ignorance of what is going on, we must solve what researchers in Artificial Intelligence call the "frame" problem by quickly constructing an operable construal of the situation. We do this by imagining or filling in details that help us decide what is most significant about the situation we face.[10] The origin of our naïve realism, then, is that we construct a construal of the situation in the absence of a much-needed check. There is no check on the confidence that our construal is correct and, in consequence, no check on our confidence that our construal will be adopted by others.

Why this absence? Why are we devoid of a mechanism to remind us that our construal is gleaned from a particular perspective and that people with other perspectives will likely construe the situation differently? We do not know. If, however, our tendency toward naïve realism is manifest in social situations, we might do well to conjoin it with the *theory of mind* theory.[11] On this view, our construal of the desires or intentions that motivate the behavior of other agents may strike us with such force that we are affectively inclined to trust it as accurate. A tendency to respond in this way may have provided anticipatory advantages during our evolutionary history; unbridled confidence in one's construal may have had greater selective value than epistemic caution. Even today, the feeling that one is right in one's assessments of others may conduce to decisive action, better learning, greater career prospects, enhanced interpersonal relations, and increased survival. Arrogance

concerning one's self may be less costly than accuracy, especially in social interactions.

The crucial upshot is that, in addition to the phenomenological quiet that accompanies our intentional actions, we naïvely overestimate the accuracy of our conscious assessments of the causes of our actions. We are seduced into thinking and feeling that the causal hypotheses that rise to conscious awareness are correct. And like the ill effects of phenomenological quiet, our naïve realism results from the constitution of our psychology. The former are by-products of a central system (motor control), whereas the latter result from an absence in the architecture of our psychology. Either way, these deficits are the direct effects of the system's normal operations; mistakes at the margins are not the issue.

The power of the directive in DP thus derives from the convergence of a range of theories, including those described here. The theory of apparent mental causation is confirmed in part by the extent to which it integrates with the forward model of motor control, the theory of naïve realism, and the theory of mind theory.[12] It thus is rational to conclude that the concept of "consciously willing" is dubious by psychological role and subject to the directive in DP. This is important for assessing the relevance of contemporary science to the view of agency presupposed in the law.

EXPLORATORY DIRECTIVES

The directive in DP is curative; it aims to cure us of conservatism and imperialism regarding dubious concepts. But it is limited. It helps us avoid what ought to be avoided without recommending an alternative strategy. The purpose of this section is to sketch a few components of an alternative that is *progressive* rather than conservative or imperialistic, and that is *exploratory* in the way that naturalists of the 19th century were explorers. Among the progressive's directives are the following:

> EH: For any capacity of the self we wish to understand, require that we frame our inquiry and our theory in terms of what is known concerning our evolutionary history.
>
> A: For any capacity of the self we wish to understand, assume that, as a consequence of our evolutionary history, it is endowed with the systemic function of anticipating objects or events relevant to organismic equilibrium, to the satisfaction of ecological demands, or to both.
>
> NC: For any conscious capacity of the human mind, expect that we will understand this capacity only after we discover the nonconscious, low-level, anticipatory mechanisms implementing that capacity.

Although EH and A appear banal, their effect on our inquiries can be substantial, altering the way we conceptualize the very capacities we wish to study. To illustrate, consider the hypothesis that human intelligence is best conceptualized as *social intelligence*. The hypothesis can be articulated in many ways, but the basic claim is that many of our affective and cognitive capacities evolved as tools enabling us to engage in myriad social relations. This is no mere speculation. It is based in part on knowledge concerning extant primate species. We know that, in one form or another, all primates are social animals,[13] capable of identifying con-specifics, recognizing social relations between con-specifics, recognizing one's own relations with others, responding appropriately to changes in those relations, and so on. And it is easy to generate hypotheses concerning the anticipatory functions of all these capacities.[14] We also know that *Homo sapiens* is the most social of primates. There is, for example, a clear difference in the breadth and depth of our cultural institutions, evidence that our capacities for social relations run wider and deeper. More specifically, human children by their fourth year clearly exercise the capacities posited in the theory of mind theory; 4 years is about the age at which most children begin to pass the false belief test. In addition, children as young as 9 months exhibit striking precursor capacities. They follow the gaze of adults, jointly focus on shared objects, imitate the behavior of others, and so on.[15] And ingenious, recent experiments suggest that human infants are reading minds, even attributing false beliefs, as early as 2 years of age.[16] By contrast, it is contentious whether other primate species possess the full suite of capacities posited in the theory of mind theory.[17]

There is also evidence of our social intelligence from neuroscience. Human infants attend preferentially to other humans. They attend to human faces more than any other visual stimuli and to human speech more than any other auditory stimuli. Infants as young as 2 days exhibit a distinctive cerebral blood flow when they hear a normal sentence but not when the sentence is played backward. And so on.[18] There is also the intriguing hypothesis that among the emotional systems implemented in the mammalian brain is what Jaak Panksepp (1998) dubs the PANIC system. This system functions to generate behavioral routines to free the organism from life-threatening situations. Effects of this system are evident in the distress calls of infants when separated from their mother, which are accompanied by physiological processes exhibited when an organism is suffocating, when it cannot catch a breath. This powerful response to separation is implemented in distinct neural structures and chemical processes identified by Panksepp. And the very same structures and processes that constitute the PANIC system also implement the reaction that adult mammals have to loss. Human grief is implemented in the brain's PANIC system.

Panksepp's hypothesis provides a striking account of our social emotions. If the basic function of the PANIC system is to generate behaviors to free the organism from threats to its life, a closely related function is to empower the organism to avoid or alleviate the experience of loss by establishing social attachments. The hypothesis is that the neural system that causes us to panic in response to loss is the very system that moves us to seek emotional attachments with others. Indeed, the PANIC system comprises neural structures known to implement certain forms of physical pain, suggesting that separation distress and grief are, literally, a form of pain and that the compulsion toward social relatedness is an anticipatory strategy for keeping some forms of pain at bay.[19]

This brief survey of the social intelligence hypothesis illustrates the power of EH and A. Notice, in particular, that our initial understanding of a capacity is altered by the application of these directives. Wegner's interpretive system, once again, is a case in point. What is the evolved, anticipatory function of a system that, by hypothesis, causes us to falsely believe that our conscious intentions cause our actions? Wegner (2007) offers several speculations, each keyed to an anticipatory, social function, and when we conceptualize our capacity for conscious willing in this way, as dedicated to some social function, our understanding is indeed altered. Instead of conceptualizing the feeling of conscious willing as evidence of a remarkable form of freedom, we conceptualize it in terms of our evolutionary history and the ways in which it prepares us for what is likely to occur next. We see the capacity as, for example, a mechanism that inclines us to inform one another about actions we are likely to perform, or a mechanism that causes us to feel a sense of obligation toward one another, and so on. And because there appears to be nothing parochial about conscious willing in this regard, the point here can be generalized: we should expect that, as our knowledge of the self progresses, our understanding of the phenomena we are trying to explain will shift in significant ways.

This shift also illustrates the power of the directive in NC. The feeling of willing, for instance, occurs at the level of conscious awareness; we are introspectively aware of some features of the process. And we tend to feel confident that the way things appear to us concerning the causes of our actions is an accurate reflection of the actual causes. Our confidence, however, is misplaced. Once we ask about the anticipatory function of any conscious capacity, we will likely discover mechanisms operating below conscious awareness that force us to revise our initial understanding. In general, it is rational to expect that our capacities for conscious experience will not be adequately understood until we discover the evolved, anticipatory functions of nonconscious mechanisms that implement those capacities.

If these exploratory directives are defensible, we must do more than withhold antecedent authority from dubious concepts. We must also fill the gaps in our conceptual repertoire left by the application of DP. We may begin by framing our inquiries with relevant knowledge from evolutionary biology and searching for the anticipatory functions of mechanisms that operate beyond the reach of conscious awareness. These are strategies informed by our best sciences of the self. And when we apply these strategies, we begin to appreciate how little of our capacities as agents we presently understand.

SKEPTICISM CONCERNING HUMAN AGENCY

If, then, the concept "conscious willing" is dubious by psychological role, and if the exploratory directives reveal that our former understanding of "conscious willing" is best replaced by a more informed understanding of the relevant capacity, then it is no longer rational to frame our inquiries in terms of this concept. It is no longer rational to assume that our alleged capacity for conscious willing is what we formerly took it to be. In particular, we cannot take it as given that our apparent capacity to consciously will our own actions reflects an actual capacity to control our actions. This is the basis for my skepticism concerning human agency.

The skeptical thesis is best formulated as an epistemic defeater: for any action we perform, we cannot justifiably claim to know from the first-person point of view the actual causes of our action. The claim is not that we never have true beliefs about the causes of our actions, but rather that we cannot reliably discriminate from the first-person point of view between cases in which our beliefs about our actions are true and cases in which they are false. This defeats the possibility of justifying, at least from the first-person perspective, beliefs about the causes of our actions. That is the lesson on which the theories described above appear to converge.

Still, this defeater appears to conflict with any number of ordinary cases in which we intuitively take ourselves to know the causes of our actions. Suppose it is Monday afternoon and you see me walking across the parking lot and entering my daughter's school. You ask me what I am doing. I tell you I am picking up Cassie and taking her to her piano lesson. Being nosy, or perhaps being an inquisitive social psychologist, you ask me why I am doing this. Being a congenial philosopher, I tell you that several weeks ago my wife and I agreed on a weekly schedule. I am taking Cassie to her lesson because that is what I agreed to do.[20]

Such intuitions, especially about cases in which the relevant action has been planned in advance, may appear to challenge the epistemic defeater.[21] But in fact they do not. In responding to your question, it is plausible to suppose that

my episodic memory quickly recollects a relevant event. It retrieves an agreement I made several weeks ago, the content of which is that every Monday afternoon during the semester I will take Cassie to her lesson. My memory did this, presumably, because the social situation demanded a timely response to your question. But the mere fact that my memory retrieved this recollection does not entail or even suggest that the content of this recollection is the actual cause of my action. My episodic memory, in recalling my earlier agreement, appears nicely attuned to considerations of relevance, but relevance requires nothing stronger than association.

Moreover, the actual causes of my action, whatever they might be, are factors that caused me to leave the house and drive to Cassie's school, and for all I can tell from the first-person point of view, the agreement I made weeks ago is causally unrelated to those factors. Of course, it certainly feels to me that the content of my recollection is causally related, but the reliability of this feeling is undermined by the theories surveyed above.[22] I also grant that I may truly believe that my recollection is causally relevant, but true belief does not suffice for justification. The theories surveyed above show that, from the first-person perspective, we cannot reliably discriminate cases in which our prior thoughts cause our actions from cases in which they do not. That is the basis of my defeater.

We might vary the locus of the objection. Instead of fixing on my present recollection of an agreement made several weeks ago, fix on the conscious thoughts that occurred just before the action. Suppose I was engrossed in work all afternoon until I happened to glance at my watch. "Oh," I exclaimed, "time to collect Cassie's music and get to her school!" Suppose I even exhorted myself: "I cannot renege on my agreement with Ann!" Suppose, finally, that upon reaching the school and hearing your question, I consciously recall and report to you the exclamations and exhortation that occurred just before initiating my action. Does this show that I know the actual causes of my action?

Not at all. It is true that my agreement with my wife was recalled to conscious awareness just before I began my Monday afternoon routine. But, again, it does not follow that this recollection is part of the actual causes of the action. All manner of nonconscious processes were no doubt occurring in me as I realized it was time to stop working, and I have conscious access to virtually none of them. And we know from the theories of Wegner and Wolpert and Blakemore and others that the things which do rise to conscious awareness often seduce us toward causal beliefs that are demonstrably false. That, to repeat, is the upshot of the theories canvased above: we know that in many cases the correlations we observe among our conscious perceptions or recollections concerning our actions are unreliable indicators of genuine causal connections, and nothing available from the first-person perspective enables us to discriminate the causes from mere correlations.

In general, it makes no difference where in the sequence we fix our attention. Even when the relevant action was planned weeks in advance, the actor is faced with an open question concerning the actual causes that finally move him to act. This is the basis of my skepticism. To refute it, we need an alternative theory of the self based upon a convergence of evidence of equal or greater strength. Short of that, no matter how unintuitive or unsettling it may be, skepticism is the rational position to adopt.

There is, moreover, an analogue to the above defeater that applies from the third-person perspective. When we claim that some other agent acted voluntarily, we posit a causal process that includes what we traditionally describe as "conscious willing." Yet, as we have seen, because the concept "conscious willing" is dubious by psychological role, it ought to be factored out of our inquiries and replaced by considerations from our evolutionary and social history. If that is right, then we cannot justifiably frame our inquiries in terms of a concept so deeply dubious. Precisely that is the basis for an additional defeater: for any action performed by another agent, we cannot justifiably claim to know that the agent acted voluntarily by consciously willing it. It must be emphasized that this defeater rests upon the above directives. The curative directive directs us to withhold antecedent authority, and the exploratory directives direct us to conceptualize the relevant capacity in terms of our history. The reason, therefore, that we cannot justifiably claim to know whether another person consciously willed her action is that, in light of our best sciences of the self, the central conceptual category has no legitimate role in contemporary inquiry.

You might worry that my skepticism refutes itself by rendering impossible all forms of rational debate. If we are indeed faced with my defeaters concerning our reasons for acting, and if adducing evidential or logical relations for a scientific or philosophical thesis qualifies as an action, it appears we can never know the reasons why anyone ever accepts one theory over another, which would undermine the very possibility of rational debate. It would seem to show, in particular, that I cannot give any reasons for skepticism concerning human agency.[23]

This worry is motivated by the apparent phenomenology of actual intellectual discussions. When in conversation you challenge some part of my view, I focus my attention on specific features of the world. When, for instance, you ask me why I hold a given thesis, I appeal to features of the world I judge to be evidentially potent. What seems crucial is that the features to which I appeal are consciously accessible to me. How could it be otherwise? How could I appeal in conversation to considerations that do not come to conscious awareness? The worry, then, is that my defeaters conflict with this bit of phenomenology. My first defeater seems to entail that I cannot justifiably

claim to know my reasons for defending the relevant thesis, which seems to preclude the possibility of my rationally defending my view. My second defeater seems to entail that you cannot justifiably claim to know my reasons for defending the thesis, which appears to make reasoned exchange between us utterly impossible.

There are several reasons why this objection is wide of the mark. I will mention just two. First, the phenomenology of our rational discussions is concerned with a relatively narrow notion of "reasons." My reasons for accepting a given thesis are patterns of evidential relations between facts in the world and the contents of the thesis, or logical relations between the thesis and other theses. As such, these sorts of reasons are limited. Even if they reveal substantive relations between the thesis and certain facts or certain other theses, they fail to explain why I endorse the thesis. It is naïve to assume that I endorse any thesis simply because of the substantive relations it bears to certain facts or to other theses. This is not to confess a foible or infirmity unique to myself. It is true of any intelligent organism whose capacities for acting are a mix of cognitive and affective capacities that operate mostly beyond the reach of conscious awareness. The evidential or logical relations I consciously acknowledge as my reasons are surely supplemented and in some instances supplanted by a host of nonconscious processes. That, at any rate, is the upshot of the scientific theories surveyed above. And that means there is a much fatter notion of "reasons" relevant to all forms of human action. This fatter notion is surely applicable to the actions we perform in the course of intellectual debates, but it is even more pertinent to actions that fall under concepts of legal responsibility. Indeed, this relatively fat notion of "reasons" is at the heart of the concept of criminal responsibility described in section I.

My second response is that it is false that my view rules out the possibility of knowing our reasons for acting. It rules out the possibility of knowing our reasons in certain ways, including ways assumed by many philosophers, legal theorists, and laypersons, but it is compatible with knowledge acquired in other ways. So long as my first defeater stands, we cannot justifiably claim to know our reasons from a first-person point of view, but that leaves open the possibility of subjecting our agential capacities, including our reason-giving capacities, to scientific investigation. My second defeater, moreover, suggests we cannot justifiably claim to know that another agent acted voluntarily by virtue of conscious willing. But that is compatible with the scientific study of our capacities as agents in terms of conceptual categories other than "conscious willing" and "voluntariness." Whether we can at present articulate an alternative concept is not to the point. It would be the most egregious form of conceptual imperialism to insist that we must preserve our traditional concepts of

agency just because we have yet to formulate other concepts informed by what is actually known about the human self.[24]

"VOLUNTARINESS" AND LEGAL RESPONSIBILITY

I come at last to the troubling implications that my defeaters raise for "legal responsibility." In the Model Penal Code, a notion of criminal responsibility is explicated in part by appeal to "voluntariness," though the explication given, as we saw in section I, is remarkably uninformative. This is so, I surmise, on the assumption that there exists sufficient shared cultural knowledge to fill the gaps in any given case. The crucial assumption, then, is that most citizens know that we are agents who sometimes determine their actions and also know when, under what conditions, our actions result from such determinations. And this crucial assumption is precisely where we meet the troubling implications of my skepticism, for this assumption is indeed false.

It should be clear by now that we do not possess shared cultural knowledge concerning the nature of human agency because we are burdened with the previously described epistemic defeaters. Thanks to progress in knowledge, we cannot justifiably claim to know from the first-person perspective the causes of our actions. Nor can we justifiably claim to know of some other person that he "determined" his own action because the central concept is so clearly dubious and our most fruitful methods direct us to conceptualize the relevant capacity in very different terms. In general, recent progress in the scientific study of the human self reveals that we do not know what kinds of agent we are. The characterization of criminal responsibility given in the Model Penal Code cannot serve its intended function.

Of course, the assumption that there exists shared knowledge of our capacities as agents runs deep and wide in our culture. That I do not deny. But the persistence and power of that assumption can be explained without granting its truth. It can be explained, in particular, by the persistence and power of concepts dubious by descent and by psychological role, and by our stubborn inclination toward conceptual conservatism and imperialism. It may be possible, moreover, to alter or eliminate this widespread assumption. If the expectations and intuitions of informed citizens are brought up to speed, if they come to reflect the larger implications of our best sciences, then appeals to commonsense may become increasingly impotent, even pathetic, when opposed to the findings of contemporary science, and our traditional concepts of the self may be revised or replaced. Until then, however, we remain burdened with laws that, for wont of knowledge of ourselves, cannot fulfill their functions.

These skeptical doubts concerning human agency are grave in two ways. They are grave because they are derived from demanding methods of inquiry. The argumentative strategy of Darwin's *Origin* is illustrative. Its power stems from the convergence upon a single view of life from a broad range of distinct phenomena concerning living things. The same holds for understanding core capacities of living things, including our capacities to deliberate, choose, and act. There is a growing convergence on the nature of the self across the relevant sciences. Not a fully articulated view of the self, to be sure, but enough to articulate some important claims: (1) A great deal of our mental lives is lived beneath the level of conscious awareness. (2) At least some of the phenomena comprising conscious awareness are partial, misleading, or illusory. (3) As a consequence, we are in a muddle about our capacities as agents; we know enough to appreciate how little we understand our experiences as selves. That is one reason why the doubts are grave: they emerge from a breadth and depth of current scientific knowledge that cannot rationally be ignored.

The doubts appear grave in another way, in terms of vital practical matters. The most pressing question is whether these doubts concerning human agency will take hold in the larger culture and, if they do, what effects they will likely provoke. One problematic effect will be the lack of a clear alternative. We live in a period of profound uncertainty about the nature of our selves; we have no choice but to endure a great deal of confusion. Another problem is knowing when to trust the converging results of human inquiry. How integrated and mature must a set of scientific theories be for us rationally to use it as the basis for policies that affect social stability, fairness, and human well-being? This is a deeply vexing question, especially for organisms who need to anticipate and feel a sense of control. Finally, there is the question of whether we have the stomach for periods of conceptual confusion, for not knowing how to think or feel ourselves as agents, and whether we will respond with creativity to better conform our beliefs and practices to the way the world actually is, or whether we will panic and revert to conservatism and imperialism.[25]

NOTES

1. We tend to attribute guilt for what a person actually does, not for what a person merely thinks. Hence the focus here on bodily movement.
2. On this point, see chapters 6 and 7 of this text.
3. The discussion in this and the next two sections is a highly compressed version of portions of my recent book (Davies 2009). Compression of course tends to distort. I hope, however, there is intelligibility enough to recommend the fuller discussion in the book.

4. Cultural mechanisms of conceptual stasis are discussed in Norris and Inglehart 2004, Richerson and Boyd 2005, chapter 6 of Davies 2009, and elsewhere. Evidence for the efficacy of psychological mechanisms of stasis is described throughout this essay and in Davies 2009.
5. In Davies 2009, part 2, I defend an additional directive to diminish the ill effects of concepts dubious by descent.
6. Since "sense" and "emotion" hardly appear equivalent, you might worry that I am being conceptually flat-footed. Perhaps so. But conceptual fussiness is a virtue only when it makes a difference in substance. Wegner is interested in not one specific affective response but rather a whole cluster. He is interested in the full range of responses to our own actions that tempt us to feel that we are conscious controllers of those actions.
7. See Wolpert at al 1995. Wolpert 1997 is an accessible overview.
8. See also Blakemore, Frith, and Wolpert 2000.
9. Choudhury and Blakemore 2006 provide a recent overview.
10. For experimental evidence, see Ross and Ward 1996, Pronin et al. 2002, Pronin et al. 2004, and Pronin 2007. This last paper surveys recent studies of the biasing effects of naïve realism.
11. Tomasello 1999 and Leslie 2000 are good overviews.
12. The full integrative picture is much broader than I can depict here. Also relevant is John Bargh's work on automaticity (e.g., Bargh et al. 2001; Dijksterhuis and Bargh 2001), Timothy Wilson's on knowing our reasons for acting (e.g., Nisbett and Wilson 1977a and 1977b; Wilson 2002), Martin Conway's on autobiographical memory (e.g., Conway and Pleydell-Pearce 2000; Conway 2003), and so on.
13. See Smuts et al. 1987 for the remarkable range of social structures among primate species.
14. Indeed, the challenge is generating experimental evidence with which to discriminate among all the possible hypotheses.
15. Tomasello 1995 and 1999.
16. Baillargeon et al. 2010.
17. Tomasello and Call 1997.
18. These references come from Cheney and Seyfarth's marvelous 2007 book on social intelligence in baboons and humans (p. 6). Cheney and Seyfarth cite Dehaene-Lambertz et al. 2002 and Peña et al. 2003.
19. See chapter 14 of Panksepp 1998. The implications of this view for the so-called "reactive attitudes" are considerable. Or so I think. See chapter 9 of Davies 2009.
20. This oversimplifies, of course. There are several reasons why I take Cassie to her lessons, including the pleasure of being with her.
21. I am grateful to Walter Sinnott-Armstrong for raising this challenge.
22. The reliability of this feeling may vary with certain features of the action and the situation in which it occurs. This appears to be an implication of the theory defended in Wilson 2002. When action and context are relatively simple and unambiguous, reliability may be greater; the interpretive system that helps us make sense of our actions may be less prone to error in uncluttered contexts. (I discuss this point in Davies 2009:146ff.) This is not, however, a difference we

can read off our feelings at the time, and it leaves ample room for error in the complex and ambiguous real-life cases in which knowledge of our reasons matters most.

23. I am grateful to Nicole Vincent for pressing this worry.
24. These two replies merely sketch the direction that a fuller response would likely take.
25. This essay descends from a presentation given at Delft Technical University in August 2009 on the occasion of an interdisciplinary conference, *Moral Responsibility: Neuroscience, Organization, and Engineering*, organized by Neelke Doorn, Jessica Nihlen Fahlquist, and Nicole Vincent. I am grateful to the organizers for a setting in which scholars from a wide range of fields engaged constructively. My travels to the Netherlands were supported by the Wendy and Emery Reves Center for International Studies at the College of William and Mary and by Silvia Tandeciaraz, Dean of Educational Policy at the College. I am grateful for both sources of support. I also received help from several good thinkers. Thanks to Stephen Morse and Walter Sinnott-Armstrong for much-needed advice; to Walter, George Harris, and Nicole Vincent for probing comments on an earlier draft; to an anonymous referee for OUP Press for constructive resistance; and to Nicole for thoughtful advice throughout.

REFERENCES

Baillargeon, R., R. M. Scott, and Z. He (2010). "False-belief understanding in infants." *Trends in Cognitive Sciences* 14(3): 110–118.

Bargh, J, P. Gollwitzer, A. Lee-Chai, K. Barndollar, and R. Troetschel (2001). "The automated will: Nonconscious activation and pursuit of behavioral goals." *Journal of Personality and Social Psychology* 81: 1014–1027.

Blakemore, S. J., C. Frith, and D. Wolpert (1999). "Spatiotemporal prediction modulates the perception of self-produced stimuli." *Journal of Cognitive Neuroscience* 11: 551–559.

Blakemore, S. J., D. Wolpert, and C. Frith (2000). "Why can't you tickle yourself?" *NeuroReport* 11(11): R11–R16.

Cheney, D., and R. Seyfarth (2007). *Baboon Metaphysics: The Evolution of Social Intelligence*. Chicago, IL, University of Chicago Press.

Chisholm, R. (1964). "Human freedom and the self." The Lindley Lecture, University of Kansas.

Choudhury, S., and S. J. Blakemore (2006). Intentions, actions, and the self. In: *Does Consciousness Cause Behavior?* S. Pockett, W. Banks, and S. Gallagher. Cambridge, MA, MIT Press.

Conway, M. (2003). "Cognitive-affective mechanisms and processes in autobiographical memory." *Memory* 11(2): 217–224.

Conway, M., and C. Pleydell-Pearce (2000). "The construction of autobiographical memories in the memory system." *Psychological Review* 107(2), 261–288.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. (A facsimile of the first edition with an introduction by Ernst Mayr, 1964.) Cambridge, MA, Harvard University Press.

Davies, P. S. (2009). *Subjects of the World: Darwin's Rhetoric and the Study of Agency in Nature*. Chicago, IL, University of Chicago Press.

Dehaene-Lambertz, G., S. Dehaene, and L. Hertz-Pannier (2002). "Functional-neuroimaging of speech perception in infants." *Science* 298: 2013–2015.

Dijksterhuis, A., and J. Bargh (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. In: *Advances in Experimental Social Psychology, Volume 33*. M. P. Zanna. San Diego, CA, Academic Press, pp. 1–40.

Kant, I. (1790). *Critique of Judgment*. (Translated by J.H. Bernard, 1951, Hafner Press.)

Leslie, A. (2000). "Theory of mind" as a mechanism of selective attention. In: *The New Cognitive Neurosciences*, 2nd edition. M. Gazzaniga. Cambridge, MA, MIT Press, pp. 1235–1247.

Nisbett, R., and T. Wilson (1977a). "Telling more than we can know: Verbal reports on mental processes." *Psychological Review* 84: 231–259.

Nisbett, R., and T. Wilson (1977b). "The halo effect: Evidence for unconscious alteration of judgments." *Journal of Personality and Social Psychology* 35: 250–256.

Norris, P., and R. Inglehart (2004). *Sacred and Secular*. New York, Cambridge University Press.

Panksepp, J. (1998). *Affective Neuroscience*. New York, Oxford University Press.

Peña, M., A. Maki, D. Kovacic, G. Dehaene-Lambertz, H. Koizumi, F. Bouquet, and J. Mehler (2003). "Sounds and silence: An optical topography study of language recognition at birth." *Proceeding of the National Academy of Science* 100: 11702–11705.

Pronin, E. (2007). "Perception and misperception of bias in human judgment." *Trends in Cognitive Science* 11: 37–43.

Pronin, E., D. Lin, and L. Ross (2002). "The bias blind spot: Perceptions of bias in self versus others." *Personality and Social Psychology Bulletin* 28: 369–381.

Pronin, E., T. Gilovich, and L. Ross (2004). "Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others." *Psychological Review* 111: 781–799.

Richerson, P., and R. Boyd (2005). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago, IL, University of Chicago Press.

Ross, L., and A. Ward (1996). Naïve realism in everyday life: Implications for social conflict and misunderstanding. In: *Values and Knowledge*. T. Brown, E. Reed, and E. Turiel. Hillsdale, NJ, Erlbaum, 103–135.

Ruse, M. (2003). *Darwin and Design: Does Evolution Have a Purpose?* Cambridge, MA, Harvard University Press.

Smuts, B. B., D. Cheney, R. Seyfarth, R. Wrangham, and T. Struhsaker (1987). *Primate Societies*. Chicago, IL, University of Chicago Press.

Tomasello, M. (1995). Joint attention as social cognition. In: *Joint Attention: Its Origin and Role in Development*. C. Moore and P. Dunham. Hillsdale, NJ, Erlbaum.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA, Harvard University Press.

Tomasello, M., and J. Call (1997). *Primate Cognition*. New York, Oxford University Press.

Wegner, D. (2002). *The Illusion of Conscious Will.* Cambridge, MA, MIT Press.

Wegner, D. (2007). Self is magic. In: *Psychology and Free Will.* J. Baer, J. C. Kaufman, and R. F. Baumeister. New York, Oxford University Press.

Wilson, T. (2002). *Strangers to ourselves: Discovering the adaptive unconscious.* Cambridge, MA, Harvard University Press.

Wolpert, D. (1997). "Computational approaches to motor control." *Trends in Cognitive Science* 1(6): 209–216.

Wolpert, D., Z. Ghahramani, and M. Jordan (1995). "An internal model for sensorimotor integration." *Science* 269: 1880–1882.