Free Will and the Bounds of the Self[1]

[Forthcoming in Robert Kane, ed. *Oxford Handbook of Free Will*. New York: Oxford University Press]

Joshua Knobe
Yale University

Shaun Nichols
University of Arizona

If you start taking courses in contemporary cognitive science, you will soon encounter a particular picture of the human mind. This picture says that the mind is a lot like a computer. Specifically, the mind is made up of certain states and certain processes. These states and processes interact, in accordance with certain general rules, to generate specific behaviors. If you want to know how those states and processes got there in the first place, the only answer is that they arose through the interaction of other states and processes, which arose from others... until, ultimately, the chain goes back to factors in our genes and our environment. Hence, one can explain human behavior just by positing a collection of mental states and psychological processes and discussing the ways in which these states and processes interact.

This picture of the mind sometimes leaves people feeling deeply uncomfortable. They find themselves thinking something like: 'If the mind actually does work like that, it seems like we could never truly be morally responsible for anything we did. After all, we would never be free to choose any behavior other than the one we actually performed. Our behaviors would just follow inevitably from certain facts about the configuration of the states and processes within us.'

Many philosophers think that this sort of discomfort is fundamentally confused or wrongheaded. They think that the confusion here can be cleared up just by saying something like: 'Wait! It doesn't make any sense to say that the interaction of these states and processes is preventing you from controlling your own life. The thing you are forgetting is that the interaction of these states and processes – this whole complex system described by cognitive science – is simply *you*. So when you learn that these states and processes control your behavior, all you are learning is that *you* are controlling your behavior. There is no reason at all to see these discoveries as a threat to your freedom or responsibility.'[2]

Philosophers may regard this argument as a powerful one, perhaps even irrefutable. Yet we doubt that people will generally find this response fully comforting. Rather, we suspect that people will continue to have the sense that if everything is controlled by these states and processes, somehow they themselves cannot be fully free or responsible.

Our aim here is to get at the sources of this discomfort and thereby gain some insight into whether or not it is warranted. We will argue that the worry people feel about these issues reflects something fundamental about the way they normally think about the sources of human

[1] We are grateful to Verity Harte, Rachana Kamtekar, Tamar Kushnir, Tania Lombrozo, Eddy Nahmias, Jonathan Phillips, and Manuel Vargas for discussion and comments on an earlier draft of this paper.

[2] Dennett (1984) provides a contemporary instance of this sort of argument, but the basic idea can be traced all the way back to Chrysippus. According to Chrysippus, my actions are produced by me precisely because they are produced by my nature and character (see, e.g., Annas, 2001, 21), so discovering that my character caused my actions could hardly count as a problem.

action. In particular, we will suggest that the worry stems from certain complex aspects of the way people ordinarily conceive of the bounds of the self.

1. Experimental Philosophy of Free Will

Consider again the picture we inherit from the sciences. It is a picture according to which human actions are caused by certain states and processes, which are in turn caused by yet earlier states and processes... and so forth. Our aim is to explain why people regard this picture as a threat to their sense of freedom and responsibility, and we will be devoting most of this chapter to developing an explanatory framework and discussing experimental studies designed to test it. But first we need to complete a preliminary step. We asserted above that people feel threatened by the prospects of a complete scientific explanation of human behavior, but we need to show that this is the case, that people actually *do* regard this picture as a threat to their sense of freedom and responsibility.

Now, one obvious way of examining people's intuitions here would be to present experimental subjects with a story of an agent who performs some dastardly deed, tell them to imagine that it was brought about through a particular sort of causal process, and then ask them whether the agent is morally responsible for what he has done. As it happens, a number of experimental studies have made use of this approach, and the results have been rather surprising. The key finding is that people show an extraordinary willingness to hold agents responsible, pretty much regardless of the nature of the process that leads up to their actions. People say that an agent can be responsible for his actions when they are told that this agent's actions are the inevitable result of his genes and environment (Nahmias, Morris, Nadelhoffer & Turner 2006), when they are told that the agent lives in a completely deterministic universe (Nahmias et al. 2006), even when they are told the agent has a neurological disorder and that if anyone else had this illness, he or she would behave in the same way (De Brigard, Mandelbaum & Ripley forthcoming).

These experimental results are fascinating, and we certainly agree that they have a lot to teach us about the nature of people's attributions of moral responsibility. But we also think that there is more to the story. It may be true that people are willing to say that an agent who performs some horrible misdeed can still be morally responsible even if his act was the result of a neurological disorder, but that does not mean that people do not see neurological disorders as being at all relevant to moral responsibility judgments. Perhaps people do see these disorders as threatening their intuitive sense of freedom and moral responsibility, but then there is some separate process at work that is overcome by the concrete, vivid, affect-laden character of the stories and therefore ends up driving people to regard these agents as responsible.

Accordingly, we conducted an experiment that made it possible to systematically vary the concreteness vs. abstractness of the questions subjects were asked (Nichols & Knobe 2007). All subjects in the experiment began by reading a description of two universes:

> Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in

> this universe was exactly the same up until John made his decision, then it had to happen that John would decide to have French Fries.
>
> Now imagine a universe (Universe B) in which almost everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it did not have to happen that Mary would decide to have French Fries. She could have decided to have something different.
>
> The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision – given the past, each decision has to happen the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision does not have to happen the way that it does.

Subjects were then randomly assigned either to the *abstract* condition or the *concrete* condition. Subjects in the abstract condition received the following question:

> In Universe A, is it possible for a person to be fully morally responsible for their actions?
>
> YES NO

Meanwhile, subjects in the concrete condition received a question that asked about a particular concrete individual who performs a specific misdeed:

> In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family.
>
> Is Bill fully morally responsible for killing his wife and children?
>
> YES NO

The results revealed a striking difference between the two conditions. Only a small minority of subjects in the abstract condition (14%) said that people in Universe A could be morally responsible, but the vast majority of subjects in the concrete condition (72%) said that Bill was morally responsible for what he had done. In other words, it seems that people do see the chain of causation in Universe A as a threat to their intuitive notions of freedom and moral responsibility, but it also seems that there is something about the presentation of a vivid, concrete, affect-laden example that drives people to think that the characters in such examples actually can be morally responsible.

Now, looking at these results, one striking feature is that participants are much more willing to ascribe moral responsibility in the concrete condition than they are in the abstract condition. A number of hypotheses have been offered to explain this effect, and researchers continue to debate the empirical and theoretical support for these contrasting opinions (Mandelbaum & Ripley 2009; Nichols & Knobe 2007; Sinnott-Armstrong 2008; Sosa 2006), but we will not be pursuing the issue further here. Instead, our aim is to focus on a simpler and more basic problem. We want to understand why it is that people are so reluctant to ascribe moral responsibility in the abstract condition in the first place. What exactly is it about Universe A that makes people reluctant to say that the agents within it can be morally responsible for their actions?

The first thing to establish here is that the results obtained thus far are not simply due to some kind of experimental artifact. For example, some may believe the response pattern arises from the fact that participants were asked whether individuals could be 'fully morally responsible.' This wording, it might be thought, appears to call for an especially metaphysically loaded sort of judgment which might differ from the sort of judgment elicited by more ordinary terms like 'free will' or 'blame.' However, subsequent studies showed that similar responses were given even when participants were directly asked whether people in a universe like the one described here could have 'free will' (Feltz, Cokely & Nadelhoffer, 2009; Roskies & Nichols, 2008) or when they were simply asked whether such people 'should still be morally blamed' (Roskies & Nichols, 2008). So it does not appear that the pattern of responses is merely an artifact of the way the question is phrased.

Others have suggested that the pattern of results might have arisen because of certain infelicities in the description of the Universe A itself. Subjects are told that "it had to happen" that the agent would act as she did, which might be taken to suggest an extreme form of fatalism according to which people's mental states and psychological processes have no impact on their behaviors (Feltz, et al., 2009; Nahmias, 2006; Nahmias, Coates & Kvaran, 2007). But this too appears to be a red herring. Subsequent work has shown that participants give the very same responses even when all of this language is removed (Misenheimer, 2009). In fact Pacer (2010) has shown that this same effect emerges when an agent's actions are explained in terms of a complex chain of cognitive processes, with each event completely causing the one after it.

Finally, some might object that the pattern of intuitions observed in these studies does not reveal any kind of general truth about human cognition but simply reflects certain idiosyncratic facts about one particular culture. Perhaps people's intuitions in these cases are influenced in some way by the contemporary American emphasis on individual autonomy. Or perhaps people have been affected by certain strands of Western philosophy or theology (arriving at certain conclusions as the result of explicit religious instruction). We certainly agree that these are very plausible hypotheses, but the empirical evidence thus far has not been kind to them. In a recent study, subjects from India, Hong Kong, Colombia and the United States were all presented with the abstract condition of the experiment described above (Sarkissian, Chatterjee, De Brigard, Knobe, Nichols & Sirker forthcoming). Strikingly, the majority of subjects in all four of these cultures said that no one could be fully morally responsible in a deterministic universe, and there were no significant cross-cultural differences in people's responses. In short, the effect here does not appear to be specific to any one culture; there really does seem to be a general, cross-cultural tendency whereby people are drawn to the view that moral responsibility is not possible in a deterministic universe.

We find these results deeply puzzling and mysterious. The question of free will is a very complex one, which philosophers have been debating for millennia. Yet ordinary people, many of whom have never thought about these questions before, seem somehow to immediately converge on one particular answer. In fact, we find this convergence even across four different cultures, with radically different religious and philosophical traditions. What could possibly explain this striking pattern of intuitions?

2. Understanding the Threat to Free Will

In thinking through this difficult question, we can begin by taking a cue from the philosophical literature. Philosophers have developed careful, systematic accounts of the ways in which a scientific perspective on human action might provide a threat to free will, and we can begin our inquiry by looking to these philosophical accounts for inspiration. Of course, it will not be possible to look in detail at each of the prominent philosophical accounts, but it seems that these accounts fall naturally into certain broad families, and we can therefore proceed by looking in a general way at a few different families of approaches.

One broad family focuses on the distinction, familiar from discussions of modern physics, between *determinism* and *indeterminism.* The distinction here centers on certain claims about the laws of nature governing our universe. So, for example, Newtonian mechanics is typically regarded as a deterministic theory, whereas certain interpretations of quantum mechanics count as indeterministic. The only way to know which type of law of nature governs our own universe is to do research in the foundations of physics.

Some philosophers have argued that questions about the laws of nature are deeply relevant to issues about free will and moral responsibility (Ginet 1990; Kane 1996; van Inwagen 1983). They say that free will and moral responsibility are not possible in a universe governed by deterministic laws. Such claims lead immediately to difficult questions of metaphysics and philosophy of science, and contemporary discussions of them focus heavily on quite complex logical principles. For example, there has been a surge of work examining the validity of the controversial 'rule beta':

$$\frac{\mathbf{N}p \qquad \mathbf{N}(p \rightarrow q)}{\mathbf{N}q}$$

(Meaning that one can infer from the claim that no one has a choice about whether *p* and the claim that no one has a choice about whether *p* entails *q* to the claim that no one has a choice about whether *q*.)

Now, it might turn out that ordinary folks can be brought to see the force of these sorts of arguments, but we think it is highly unlikely that arguments from this family are getting at the root of people's intuitive worry about free will. In particular, it seems that people do not ordinarily understand their world in terms of laws of nature. They might acquire the concept of laws of nature when taking physics courses, but it seems unlikely that this concept plays an important role in their intuitive conception of how the world works, and it therefore seems implausible that their principal worry about free will is a worry that the laws of nature might turn out to be deterministic. The intuitive problem presumably lies elsewhere. (Indeed, even within the traditional philosophical literature, the free will problem was often posed not in terms of deterministic laws of nature but simply in terms of the idea that each event might be caused by some event that occurred before it; see, e.g., Spinoza 1992/1677.)

Let us therefore turn to the second major family of philosophical accounts. This second family focuses on the *self* and the worry that the self might turn out not to be the source of human action (Nietzsche 1989/1887; Spinoza 1992/1677; Strawson 1986). On this second sort of view, the central worry is not really about deterministic laws of nature. Rather, determinism is merely serving to crystallize or make salient another sort of worry. The central concern is that we might discover that when an agent acts, *she herself* is not in some relevant sense the source of her own actions.

In a series of recent papers, Eddy Nahmias and colleagues have argued that people's intuitive worry about free will actually takes this second form (Nahmias 2006; Nahmias, Kvaran & Coates 2007; Nahmias & Murray forthcoming). Their suggestion is that the intuitive worry about free will stems from the thought that the causal chain leading up to our actions might turn out to bypass the self entirely. In other words, the worry is that the self is *epiphenomenal* with respect to action.

Perhaps the easiest way to bring out the force of this idea is by introducing a simple example. Suppose we discover that John's actions are entirely determined by the states of his brain. We might then experience a worry that John does not have free will. But why? On the hypothesis under discussion now, the worry here is fundamentally about the role of John's self in his own actions. That is, when people hear that John's actions are determined by his brain states, they do not think: 'Oh no! So our universe is governed by deterministic laws that link brain states to behavior...' Rather, they think something like: 'Oh no! So it isn't really *John* who gets to decide what to do; it's merely his brain that is controlling all his actions...' Nahmias and colleagues have presented an impressive array of experimental evidence for this hypothesis, and we think that they are on exactly the right track.

But, of course, even if this hypothesis helps to answer certain questions, it also raises a host of new questions of its own. Why exactly would anyone worry that the self is epiphenomenal in this way? Why would our experimental stimuli trigger that worry? And what is it about contemporary work in cognitive science that makes the worry seem so pressing?

One way to address this question would be to suggest that people are simply falling victim to some kind of straightforward confusion. One might think that people are somehow failing to read the vignettes in the questionnaires correctly. Or that they are getting confused about the relationship between brain and mind. Or that they don't quite understand what determinism involves. All of these hypotheses are plausible ones, which would be worthy of further theoretical and empirical exploration.

Our aim here, however, is to propose a very different hypothesis. We want to suggest that people's intuitions in these cases are not merely the result of confusion but reflect something deep and fundamental about the concepts they ordinarily use to make sense of the world. In particular, we will argue that these intuitions are pointing at something important about the way people ordinarily think about of the self.

3. Three Conceptions of the Self

We noted above that the pattern of responses in recent experimental studies leaves us with a puzzle. Most people have presumably given little thought to the problem of free will, and yet, when experimental philosophers present them with these strange questions about alternate universes, they seem somehow to converge on the same pattern of responses. Given that the questions are so bizarre and unfamiliar, why is it that most people respond in this same way? We can now propose a hypothesis about how this convergence arises. Our hypothesis will be that people arrive at the same intuitions about free will because they share the same basic way of understanding the self.

To really unpack this hypothesis, we will report a series of new experimental studies. But first we need to refine our conceptual framework. We begin by considering a basic question – what exactly is the self? More specifically, what are the bounds of the self – what falls inside and outside the self?

The issue here seems straightforward enough. Suppose that we are observing John and trying to figure out whether he himself is in control of his actions. To do this, we need to draw a distinction between two different types of factors. On one hand, there is John himself; on the other, there is the broader situation in which he happens to be embedded. But how exactly can we distinguish between these two types of factors? In some cases, it may all seem perfectly simple – the temperature in the room is clearly an aspect of John's situation, not a part of John himself – but there may be other cases in which the distinction proves harder to grasp. If John has a broken arm, would that be a problem in John himself or merely a difficult aspect of the situation that he happens to be confronting? What if he had a brain tumor? Our aim here is to arrive at a better understanding of the way people ordinarily make sense of these questions.

Fortunately, we already have before us a rich source of hypotheses. After all, in philosophy, questions about the nature of the self are at least as old as questions about free will and determinism. So we might begin by considering various conceptions of the self that have been articulated by philosophers. We focus here on three particularly prominent approaches.

3.1. The bodily conception of the self

One conception of the self is that the self contains everything from the skin in. So your brain is part of you, but so are your feet, your intestines, and so forth. This conception certainly does have a strong appeal. The body is, after all, the primary means by which we typically identify each other. And if a falling tree breaks John's leg, it seems that this damages *John* – it would seem implausible to say that the tree didn't hit John himself but only his body.

In philosophical work on the self, the identification of the self with the body forms a venerable tradition. It emerges, for example, in Nietzsche's dictum:

> "Body am I, and soul"—so says the child. And why should one not speak like children?
>
> But the awakened one, the knowing one, says: "Body am I entirely, and nothing more; and soul is only the name of something in the body."
>
> […] Behind your thoughts and feelings, my brother, there is a mighty lord, an unknown sage—it is called Self. It dwells in your body, it *is* your body. (Nietzsche 1999/1883 I 4, our translation)

This basic conception of the self has been developed, though with important variations, in much contemporary work within the analytic tradition (e.g., Carter 1990, Olson 1997, Williams 1970). Such work emphasizes that human beings are fundamentally animals and that a proper understanding of the self must take account of our nature as embodied organisms.

If the bodily conception is right, it's hard to see how the epiphenomenal worry would even get off the ground. Certainly scientific work provides no basis for worrying that bodily processes are left out of decision-making. Although there is abundant scientific disagreement about the self and its role in decision, none of the prominent accounts would deny that decisions are generated by stuff inside our bodies.

But the fact that the bodily conception of the self renders the epiphenomenal worry toothless does not end the discussion. For the bodily conception of the self is hardly the only available conception. On a 'thinner' conception of the self, certain factors that lie within the body could nonetheless fall outside the bounds of the self. On such an account, even if we acknowledge that an agent's actions are under the control of factors within that agent's body, this doesn't settle the question as to whether those actions were under the control of the agent herself.

Our aim is to explore the idea that people's ordinary understanding of the self might actually rely on such a thinner conception. But for present purposes, we are focusing on *philosophical* treatments of the self. And the bodily conception of the self has been disputed from the beginning of Western philosophy. Already in the Socratic dialogues we find a vigorous rejection of the bodily view.[3] As Socrates is about to drink his hemlock, Crito asks, "in what fashion are we to bury you?" Socrates then upbraids Crito for thinking that the object that will be buried is Socrates himself. The object that will be buried, he says, is merely a *body*, while Socrates himself is something quite different:

> Friends, I can't persuade Crito that I am Socrates here, the one who is now conversing and arranging each of these things being discussed, but he imagines I'm that dead body he'll see in a little while, so he goes and asks how he's to bury me!

After Socrates dies, there will still be a body in the room, but Socrates himself will no longer be present. Therefore, the body of Socrates and Socrates himself must somehow be distinct things.

This is a powerful philosophical argument. But if the self isn't simply the body, what other conceptions are available? Two prominent conceptions follow.

3.2. The psychological conception of the self

Instead of a bodily approach to the self, we might adopt a more restrictive notion of the self – one on which fewer things count as part of the self and more counts as external and merely part of the environment. One might adopt the view that only *psychological* things associated with a body are part of the self; the feet, intestines, and so forth are merely external objects to which the self happens to be attached. What really matter are the memories, convictions, aspirations, etc. That's what constitutes the self. On this view, the physical features of one's body are often obstacles to the self. For instance, the physical features can impede the aspirations and convictions that make the self. If my foot is broken, this is plausibly a problem that I face. My broken foot doesn't constitute me, even in part; rather my broken foot is external to who I really am – it's a problem that I confront.

This psychological approach to the self also has much in its favor. It draws support from the philosophical tradition according to which much of what we regard as most important about our selves is precisely our memories, convictions, and so forth (cf. Locke 1979/1847, Parfit 1986). It also draws support from the assumption, widely shared within cognitive science, that the only tenable view of the self will be given in terms of psychological states and processes.

[3] Strikingly, the bodily conception is also disputed in the independently developed philosophical tradition of India. In the *Chandogya Upanisad*, the sage Maghavan says "This body… is mortal… So, it is the abode of this immortal and nonbodily self." (6.12.1).

If this psychological conception is the right view of the self, it's easier to see how science might show that the self isn't in charge. For if science shows that out behavior is caused by non-psychological processes in our bodies, this will mean that the behavior isn't caused by the self. On the psychological conception of the self, if the self is the cause of our behavior, it must be the case that our psychological features cause our behavior. If instead science shows that psychological features are irrelevant to behavior, then the self is indeed epiphenomenal.[4]

Notice, however, that on a view like this one, the purely cognitive sciences would be no threat at all to free will. Learning that one's actions are caused by one's own mental states would only help to confirm the conviction that one was free.[5]

3.3. The executive conception of the self

There is, however, a third possible conception of the self on which the cognitive sciences pose a deep and abiding threat. Instead of adopting the view that the self is just a bunch of mental states, one might suppose that the self is really some further thing, something over and above the various mental states one might have. On this view, the particular mental states you have are external to the self, much as intestines are external to the self on the previous view.

It's easy to see advantages of the view. Just as my broken foot is plausibly external to who I am, there is some force to the idea that the particular psychological characteristics I have are external to who I am. Thus, suppose John has always had a longing to become a famous guitarist but also suffers from extreme stagefright. On the conception of the self under discussion here, both the longing for fame and the stagefright would just be aspects of the situation John happens to be confronting. However, there would then be some further thing – John himself – which could consider these various drives and emotions and make a decision based on them. So John would count as in control of his actions to the extent that his actions were determined not by his individual psychological states (the longing, the fear, etc.) but by an executive that could consider these states and arrive at a decision.

This view of the self has deep roots in intellectual history. It is plausibly the dominant strand of thought about the self in ancient philosophy. There, the common view is that the self is the soul, the seat of psychological states and the source of action. This is particularly clear in Stoic philosophy, in which the soul is a *commanding-faculty* ("hegemonikon"), which thinks, plans, and decides (Baltzly 2008), and it is this commanding faculty that is thought to be separable from the body (cf. Sextus Empiricus 1949/2000 7.234).

We thus seem to get a glimpse of the executive conception of the self in ancient philosophy, but it is only in the early modern period that this executive conception is explicitly differentiated from the psychological conception. For instance, on Reid's theory, the self is the

[4] If, as suggested by the hypothesis of folk dualism (e.g. Bloom 2004, 2006), people think of the mind as something entirely non-physical, separate in every way from the human body, then we should expect people to find this threat quite vivid. Demonstrating that our behavior is controlled by physical processes (neurons or whatever) would undermine the idea that it is one's (non-physical) mental states that are doing the work.

[5] Obviously there are further distinctions available within the general psychological approach to the self. For instance, one might think that the self is a proper subset of one's psychological states. Researchers have proposed a number of accounts along these basic lines, each picking out a different subset to count as the self. (Frankfurt 1971, Watson 1975, and Wolf 1990 offer somewhat different accounts. For discussion of a *very* different proposal about which psychological states constitute the self, see Gide 2000/1902.) In the experimental studies reported below, we will be asking whether ordinary people hold a simple psychological conception of the self, but these same experimental methods could be used to ask whether people's intuitions follow any version of the subset view.

soul, and it is also the *agent* that causes decisions.[6] Reid memorably insists that the psychological approach to the self perverts the relationship between the self and its psychological states: “I am not thought, I am not action, I am not feeling; I am something that thinks, and acts, and suffers.” (Reid 1785/1969 341). On such a view, the psychological states don’t constitute the self, they *belong* to the self, and the self makes its decisions *in light of* the psychological states, but not as a simple consequence of the states.

If this view of the self is right, then an account of decision making in terms of psychological churning and processing would leave the self out entirely. If the self is something other than our psychological characteristics, then insofar as psychological states are in the driver’s seat the self isn’t. Of course, just as you have intestines, you also have various desires, emotions, etc., but on this view of the self, it is not as though these desires, emotions, etc. just interact with each other in some complex way and then produce your actions. Rather, you are confronted with these desires, emotions, etc., and then *you* choose in light of all of them which action to perform. On this view, the self is an *executive* that stands apart from the particular mental states that inform her decision. In this sense, it is like the president of a country. The president might listen to various advisors representing various constituencies and then make a decision. But the president himself is not just a bunch of advisors and constituencies. He is some further thing that can listen to the advisors – but can also choose to go against their advice.

3.4. Commonsense and the three conceptions

On some of these views of the self, it is simply obvious that if your actions are controlled by your desires and values, then they are controlled by you, whereas on other views, it should be obvious that if your actions are controlled by your desires and values, then they are not controlled by you. At first glance, it might seem that our task is correspondingly clear – if we want to determine whether epiphenomenalism is a looming threat to our ordinary view of the self in action, we need only discover which notion of the self is at play in common sense.

We fear, however, that the task is not nearly so straightforward. For we think that the diverse philosophical views on the self are not pristine inventions of academic philosophers. Rather, we suspect that these different views of the self all reflect important strands of commonsense thought about the self. In some contexts, people think of the self as the body, in other contexts, it’s the psychology, and in other contexts, it’s neither. People shift between these different views of the self depending on the way in which they are thinking of the problem. A proper theory in this domain needs to adequately reflect this complexity, getting at the sources of our attraction to the different conceptions and the nature of the ensuing conflict. But merely speculating about this is one thing. What we need to do now is formulate hypotheses about what factors of a situation might make people incline to one conception rather than another.

4. Shifting perspectives and the conceptions of self

Developing hypotheses about the factors influencing how people think about the self difficult task, and it might be best to begin by approaching it somewhat indirectly. Instead of starting in immediately with these vexed questions about the constituents of the self, we can begin by looking at a far simpler analogue: a question about the constituents of a *corporation*.

[6] On such "agent-causation" views, the agent causes the action without the agent herself being caused to do so. In that sense, agent-causal views do reject deterministic accounts of decision making. For on agent-causal views, the agent herself is not determined to decide one way rather than another.

Looking at a typical business deal, one might observe that there is a corporation which, taken as a whole, is performing certain actions, and one might also see that there are various other objects – buildings, employees, etc. – which appear to be playing an important role. But now it seems that a question arises about the relations among these distinct entities. Are the buildings, employees, etc. literally constituents of the corporation itself, or should one say that the corporation is some radically different kind of thing, such that things like buildings and employees could never literally be parts of it? We maintain that people do not have any single, stable approach to making sense of this sort of question. They have a capacity for thinking about things like corporations, and they have a capacity for thinking about things like buildings and employees, but they do not have a single fixed picture of the relationship between the former and the latter. Instead, their intuitions about questions like this one can vary greatly depending on the particular type of perspective they adopt.

First, suppose that we are looking out at an entire city and thinking about what might be going on in each of its various neighborhoods. One of us might point over at a particular location and say: 'Those buildings are part of Microsoft, while those over there are part of Intel.' In this sort of context, such a remark might seem perfectly natural. We could immediately see how certain buildings would be part of one corporation while others could be part of a separate corporation.

But now suppose we adopt a different perspective. Suppose we zoom in on one specific building and consider it in detail. We are thinking about the building's physical structure, the chemical composition of its bricks and mortar. It might then seem a bit bizarre to suppose that this building – an actual physical object – could literally be a part of a corporation. Indeed, it may begin to seem that if one really understood what it is to be a building and what it is to be a corporation, one would have to see that the former just isn't even the sort of thing that could possibly be a part of the latter.

Assuming now that these claims about people's intuitions are correct, let us sum up the basic pattern. There seems to be a distinction between the perspective we adopt when we are *zooming out* and the perspective we adopt when *zooming in*. When one 'zooms out' to consider a vast panorama of different objects and processes, it may seem obvious that certain buildings count as part of a corporation. But if one then 'zooms in' to think about one particular building in detail, it may begin to appear that this building could not possibly be part of a corporation at all. Hence, it may be that people do not have any single, stable view about what lies inside or outside of a corporation. As their perspective changes, so does their conception of the corporation's constituents.

We now want to suggest that a similar phenomenon arises for people's conceptions of the self. That is, we will be arguing that people do not have any single, stable view about what lies inside or outside of the self. People have a capacity for thinking about agents, and they have capacities for thinking about things like bodily parts and mental states, but they do not have a single fixed sense of the relationship between the agents themselves and the bodies and mental states with which they are associated. Instead, people's intuitions about this relationship depend in part on the perspective they are adopting at the time.

Consider in this light our earlier question about the relationship between the body and the self. Do the parts of John's body truly count as parts of John himself? Or are they merely aspects of the situation in which he happens to be embedded? One way to approach this question would be to consider the entire planet and to ask, for each thing on this planet, whether it lies inside or outside of John. When we zoom out this far, it may begin to seem perfectly obvious that anything in John's body has to count as a genuine part of John. (It would seem a bit bizarre in this context to say: 'John's pancreas is not actually a part of *him*; it is just a part of his body, which is something else entirely.') But now suppose we adopt a different perspective. Suppose we zoom in closely and begin thinking in detail about all the factors, both mental and physical, that went into one particular decision John made on one specific occasion. John has always spent most of his income on his children but is now facing some serious health problems, and the question is whether he will decide to start spending his money on expensive medicines instead. Looking at this sort of case, our intuitions may begin to shift. It may begin to seem quite tempting to suggest that John's pancreas is not best understood as a part of his self at all – that it is merely an aspect of the situation in which *he* happens to be embedded.

Or consider the question about the relationship between a person's self and his or her various mental states. Suppose first that we are confronted with some quite intricate social situation, involving dozens of different people, and we are trying to get at the source of a particular problem, say, that a business venture is underfunded. If we now discover that the problem can be traced back to something about John's anxieties and fears, we might immediately conclude that the problem lies within John himself. (It would be a bit odd even to consider the objection that John's emotions are not properly regarded as parts of John.) But now suppose we switch over to a more zoomed-in perspective. Suppose we focus on the details of the case in which John is forced to make a decision. He plans to perform a specific action but then finds himself overcome by anxieties and fears, and the big question is whether he will be able to go through with the plan nonetheless. In this latter sort of context, it might seem fairly natural to regard the anxieties and fears, not as parts of John himself, but as a particularly difficult aspect of the situation he now faces. John, when we focus closely on him, is identified with a thin, executive self.

With this general theoretical framework in place, we can now put forward a specific testable prediction. Consider a case in which an outcome is caused by John's emotions (without involving any actual choice on his part), and now suppose we ask whether John caused the outcome. People should respond differently to this question depending on their perspective:

- When they zoom out to consider the broader context, they will regard John's emotions as part of John's self, and they will therefore conclude that John himself *did* cause the outcome.
- But when they zoom in to consider that behavior in isolation, they will adopt a conception according to which John's emotions do not count as a part of his self, and they will conclude that he *did not* cause the outcome.

This, at least, is the theory. To decide whether this theory is true, we need to gather some additional data.

5. Experimental Studies

The basic methodology behind our experimental studies is a simple one. Subjects were presented with cases in which it was clear that John's body or his psychological states brought about a particular outcome, and they were then asked whether they agreed with the claim that

John brought about that outcome. The aim was to use this methodology to get a better understanding of how people think about the bounds of the self.

If people think that John simply is his body or his psychological states, then whenever John's body or his psychological states causes an outcome, they should think that John himself caused that outcome. But that is not the pattern of intuitions we predict. Instead, we predict that people's intuitions will vary depending on their perspective. The more they zoom out to consider a broader context, the more they should feel that John counts as the cause of the outcome in question. By contrast, the more they zoom in to consider the details of the process leading up to this one particular behavior, the more they should feel that John himself does not count as a cause at all.

Study 1

Ultimately, our aim is to look at the contrast between zoomed-out cases and zoomed-in cases, but before introducing zoomed-out cases, we thought it might be best to explore how people think about the issue in zoomed-in cases in which we ask people to look in detail at the process leading up to one specific behavior. In particular, we wanted to determine whether there are conditions under which people would say that the person *didn't* cause an outcome, even though the outcome was a product of the person's psychological states.

So, for this first experiment, we looked at a case in which John's eye blinks rapidly. Each subject was randomly assigned either to one of two conditions. In one condition, subjects received what we will call the *choice-cause* case:

> Suppose John’s eye blinks rapidly because he wants to send a signal to a friend across the room.
>
> Please tell us whether you agree or disagree with the following statement:
>
> - John caused his eye to blink.

In the other condition, subjects received what we will call the *emotion-cause* case:

> Suppose John’s eye blinks rapidly because he is so startled and upset.
>
> Please tell us whether you agree or disagree with the following statement:
>
> - John caused his eye to blink.

Subjects rated each statement on a scale from 1 ('disagree') to 7 ('agree').

Notice the force of the question in the emotion-cause case. The scenario makes it clear that the blinking is being caused by John's psychological states, namely, by his being upset. So if participants actually do hold a psychological conception of the self (according to which John simply *is* his psychological states), they should conclude that the blinking is caused by John himself.

But we predict that participants will *not* adopt a psychological conception of the self in this case. Since the scenario encourages participants to zoom in on the processes underlying a specific behavior, they should instead adopt a narrower conception, like the *executive* conception. On this conception, John's emotions do not actually count as part of John himself.

Instead, John is seen as some kind of further entity which can take into account these various emotions and then make a choice. Thus, participants should regard John himself as a cause in the choice-cause condition but not in the emotion-cause condition.

Indeed, that is exactly what we found. Subjects tended to agree with the claim that John caused the outcome in the choice-cause case, while they tended to disagree with the claim that John caused the outcome in the emotion-cause case. This difference was statistically significant.[7]

Our aim is to use results like this one as part of an extremely simple argument. It is clear in a case like this one that the ensemble of John's psychological states caused an outcome, but people nonetheless say that John himself did not cause the outcome. Therefore, people do not conceive of John himself as simply being the ensemble of his psychological states.

Of course, this isn't the only hypothesis that fits the data. We won't be able to address all available alternative hypotheses here, but we can chip away at some of the major contenders.

Study 2

To begin with, we are assuming that the above case is a clear instance of John's psychological states causing his behavior. But someone might reject this assumption. Someone might say that people's ordinary concept of causation is actually more complex than we have been assuming and that people's ordinary intuition is that these behaviors are neither caused by John nor caused by his psychological states. Alternatively, someone might say that reactions of startle and anxiety aren’t the kind of psychological states that constitute a person. (It might be said, e.g., that only states like thoughts and goals are truly internal to the self.) These are certainly reasonable worries, and to address them, we conducted a second study.

In this second study, all subjects were given the following case:

- John’s hand trembled because he thought about asking his boss for a promotion.

All subjects were then asked whether they agreed or disagreed with two statements about this case. One of the statements was quite similar to the one used in the emotion-cause condition of Study 1:

- John caused his hand to tremble.

The other statement was then exactly like the first, except that the word 'John' was replaced with 'John's thoughts':

- John's thoughts caused his hand to tremble.

The order of the two statements was counterbalanced, but there were no order effects.

As predicted, these two statements led to two very different responses. People tended to disagree with the statement that John caused his hand to tremble, but they tended to agree with the statement that *John's thoughts* caused his hand to tremble.[8] In other words, the results yielded an especially stark version of the effect obtained in the earlier study. People are apparently happy to say that the outcome *was not* caused by John himself but *was* caused by John’s thoughts. This suggests that people conceive of John himself as being something distinct from his thoughts.

[7] *N* = 30 people spending time in a New York public park, mean for choice-cause 5.5 out of 7, mean for emotion-cause 2.6 out of 7, *t*(28) = 3.8, *p* = .001.

[8] *N*=41 students in introductory philosophy classes at University of Arizona. Mean agreement response for "John's thoughts caused" was 5.8 out of 7; mean response for "John caused" was 3.76 out of 7. This difference is statistically significant (*t*(40)= 3.97, *p*<.001)

Moreover, we expect that much the same result would emerge if we picked any other class of mental states – desires, urges, convictions, etc. That is, we expect that people would be willing to say about various cases that John's *desires* (convictions, urges, etc.) caused his behavior even though *John* didn't cause it. If the result generalizes across mental state types, then this suggests that, at least in these contexts, people reject the psychological conception of the self, apparently in favor of the executive conception.

We recognize, however, that there might be other ways of explaining these data. For example, one might seek to explain people's intuitions in these cases by positing a fairly simple conception of the self and then accounting for all the puzzling results by assuming that people have a quite complex concept of causation. (The idea then might be that people's concept of causation is somehow sensitive to the distinction between acts that were performed on purpose and those that were not; see, e.g., Lombrozo 2009.) Or perhaps there is some third form of explanation available here, one that relies neither on people's concept of the self nor on their concept of causation. Just looking at the experimental results we have presented thus far, it seems that numerous possible approaches might prove viable here.

Study 3

Although several approaches might explain the above results, our theory also generates another, very different prediction that does not fall naturally out of any of the other possible approaches. As we have seen, the theory says that when people zoom in to consider one specific behavior (like blinking), they will tend to regard the agent's body and psychological states as falling outside of the self. But the theory also says something further. It says that when people zoom out to consider a broader context, they will change their conception and adopt a view according to which the body and mental states actually *are* part of the self. Hence, if we can just get subjects to zoom out a little more, they should end up concluding that John himself actually *is* the cause of all the outcomes that are caused by his body or mental states.

For an illustration of the basic idea here, consider the following vignette:

> Suppose that John has a disease in the nerves of his arm. He experiences a sudden spasm, his arm twitches, and his hand ends up pushing a glass off the table. As the glass strikes the floor, there is a loud crashing noise.

In this vignette, John has a medical condition that leads to a spasm. Will people think that this condition is a part of John himself? Well, their answer should depend on the perspective they adopt. To they extent that they zoom in and think in detail about the processes leading up to that one arm movement, they should regard the medical condition as falling outside the self. But suppose we force them to change perspective. Suppose we get them to think about the broader context, including not only the bodily movement but also the table, the glass, the crashing noise, and so forth. If the theory is correct, they should then change their conception and adopt a view according to which the medical condition actually is part of the self. Hence, they should then be drawn to the idea that anything caused by the medical condition actually was caused by John.

To test this hypothesis, we conducted a further experiment. All subjects were given the vignette about the spasm that pushes the glass off the table. Subjects were then randomly assigned either to the *zoomed-in* condition or to the *zoomed-out* condition. Subjects in the zoomed-in condition were asked whether they agreed or disagreed with the sentence:

- John caused his arm to twitch.

This sentence was designed to make subjects focus in on the details of the process leading up to one particular bodily motion, and according to the theory, it should therefore lead them to adopt a 'thin' account of the self whereby John's bodily parts and even his mental states do not count as falling within John' self.

Subjects in the *zoomed-out* condition were asked whether they agreed or disagreed with the sentence:

- John caused the loud noise.

This sentence was designed to make subjects think more broadly about the whole situation in which John was embedded, and it should therefore lead them to adopt a 'thicker' notion of the self, according to which the nerves in John's arm count as a part of John himself.

As predicted, these two sentences led to two very different patterns of intuition. Subjects in the zoomed-in condition tended to disagree with the claim that John caused his arm to twitch, whereas subjects in the zoomed-out condition tended to *agree* with the claim that John caused the loud noise. This difference was statistically significant.[9]

Notice the puzzling character of people's intuitions here. The noise was clearly brought about through the twitching of the arm, yet people somehow conclude that John *did* cause the noise but *didn't* cause the twitching. It seems difficult to make sense of this asymmetry by supposing that there exists some single object – John – which people regard as the cause of the noise but not of the twitching. Rather, the natural explanation here would be that people are adopting different conceptions of the self in the different cases. They adopt a thinner conception in the zoomed-in case, a thicker one in the zoomed-out case. Then the thicker their conception of the self, the more inclined they are to regard the self as a cause of the series of events that unfolded.

Study 4

Thus far, we have been looking separately at a number of different variables that affect people's intuitions. First we looked at the distinction between different types of actions (choice-cause vs. emotion-cause); then we looked at the distinction between different types of perspectives (zoomed-in vs. zoomed-out). The experimental results seemed to indicate that both of these variables had an impact on people's intuitions. For this final study on zooming, therefore, we wanted to conduct a single experiment that would allow us to systematically examine all possible combinations of these two variables.

In other words, we wanted to look at people's intuitions about the four possible cases in the following 2 x 2 table:

	Zoomed- In	**Zoomed-Out**
Choice-Cause	x	x
Emotion-Cause	x	x

[9] *N*= 40 people spending time in a New York public park. The mean rating for the zoomed-in condition was 2.0 out of 7; the mean rating for the zoomed-out condition was 4.8 out of 7,$t(38) = 4.6$, $p < .001$.

Each subject was assigned to receive one of these four cases. By looking at the intuitions subjects had in each case, we hoped to get a better sense for the impact of the two variables under discussion thus far.

For the zoomed-in cases, the set up went as follows. In the choice-cause condition, participants were given the following instructions:

> Imagine you just observed the following:
>
> A bee lands next to John and his hand withdraws.
>
> Now suppose you learn that John's hand withdrew because he is afraid of bees.
>
> Please tell us whether you agree or disagree with the following statement:
>
> - *John caused his hand to move.*

In the emotion-cause condition, the case was exactly the same, except that the word 'withdraws' was replaced with 'trembles':

> Imagine you just observed the following:
>
> A bee lands next to John and his hand trembles.
>
> Now suppose you learn that John's hand trembled because he is afraid of bees.
>
> Please tell us whether you agree or disagree with the following statement:
>
> - *John caused his hand to move.*

Hence, John performs exactly the same behavior in the two cases; the only difference is that in the choice-cause condition he does so presumably as a result of a choice, while in the emotion-cause condition his behavior seems to be directly caused by his emotions.

For the zoomed-out cases, the set up was exactly parallel except that John's movement (withdrawal or trembling) knocks over a glass of milk. Subjects were then asked whether they agreed or disagreed with the statement:

> - *John caused the milk to spill.*

Subjects rated each of these sentences on a scale from 1 ('disagree') to 7 ('agree').

The theory under discussion here predicts that people should respond differently in the zoomed in cases, but they should not respond differently in the zoomed out cases. For we maintain that when people zoom out, they are more promiscuous about what counts as part of the self; by contrast, when they zoom in, they tend to think of the self as executive. The results confirmed this prediction by revealing a striking difference between the zoomed-in and zoomed-out cases.

	Zoomed- In	**Zoomed-Out**
Choice-Cause	6.10	4.95

Emotion-Cause	3.95	5.48

In the zoomed-in cases, people regarded John as a cause when he chose to move his hand but not when his behavior was produced directly by his emotions. In the zoomed-out cases, by contrast, there was no such difference: John was regarded as a full-fledged cause either way.[10] This fits perfectly with the proposal that in the zoomed-out condition, people are willing to take a very thick view of what counts as the self, but in the zoomed-in case, people are inclined to think of the self as a thin executive.

Of course, we are open in principle to the idea that this asymmetry in people's intuitions could be explained without positing a shift in their conceptions of the self. However, we have not been able to come up with any alternative hypothesis that can explain the full pattern of intuitions revealed in these studies. At least for the moment, then, we will be proceeding on the assumption that people's conception of the self actually does shift depending on the perspective they employ.

6. The executive self and cognitive science

We began with a puzzle about the basic picture of the mind coming out of cognitive science. This picture says that human actions are caused by certain psychological states and cognitive processes, which are in turn caused by other states and processes, and so on, back into the past. Such a view might seem relatively harmless – perhaps even obviously true – but recent research indicates that people often find it strikingly unsettling. They appear to regard it as a serious threat to the possibility of human free will.

A proper explanation of people's intuitions here should allow us to see why the picture coming out of cognitive science leaves them with this feeling of unease, but it should also help us to understand why the issue is so characteristically *confusing*. It should help us to see why people so often feel pulled in competing directions, why they come to think that there is some kind of deep philosophical problem here that needs resolving.

Our aim now is to take the theoretical framework we have been developing thus far and use it to address these questions. We proceed in two steps. First we provide experimental evidence that the picture coming out of cognitive science goes against people's ordinary understanding of human action. Then we argue that it is this departure from people's ordinary understanding that generates the perceived threat to free will.

6.1. The self and cognitive science

Researchers in cognitive science often rely on an analogy between the mind and a piece of computer software. In a typical piece of computer software, one finds certain lines of code and certain data structures, and everything the computer does can be understood in terms of the

[10] *N*=41 students in introductory philosophy class. The data were analyzed using a 2 x 2 ANOVA, with zoom (zoomed-in vs. zoomed-out) as a between subject factor and action type (choice-cause vs. emotion-cause) as a within-subject factor. The results showed no main effect of zoom, $F(1, 39) = .13$, $p = .7$, though there was a main effect of action type, $F(1, 39) = 5.5$, $p < .05$. Most importantly, there was a significant interaction effect, $F(1, 39) = 14.9$, $p < .001$, indicating that the impact of action type is larger for the zoomed-in case than for the zoomed-out one.

operations of the code on the data. The dominant view in cognitive science is that the mind works in more or less the same way. In place of lines of code and data structures, we have cognitive processes and mental states, but the basic explanatory paradigm remains the same: one posits certain operations of the processes on the states, and these operations are supposed to determine everything we ever think or do.

On the account we have been developing here, people's ordinary understanding does not consist in some kind of coherent theoretical viewpoint but rather involves applying different conceptions in different sorts of cases. Most importantly for present purposes, when people adopt a more zoomed-in perspective, we suggested that they end up with a conception that is really quite different from the standard cognitive science picture. In this perspective, they do not conceive of the mind as being fundamentally like a computer. They do not think that human behavior is just a product of cognitive processes operating on mental states. Instead, they adopt a conception according to which there exists some further thing – the self – that stands outside all of these states and processes and can choose whether to obey them or not.

To more directly explore these questions, we conducted one final experiment. Each participant received a scenario about a computer and story about a human being. (The order of scenarios was counterbalanced.) The computer scenario stipulated that all of the computer's programming was directing it against a particular behavior. We then asked whether the computer might cause that behavior nonetheless. The case went as follows:

> VQ5T is a computer that has a robotic hand. The robotic hand is positioned next to the power button for a device that is delivering electrical shock to a rat in an experiment. If VQ5T moves its hand to the right, it will push the button and stop the shocks.
>
> VQ5T has the information that if it moves its hand to the right it will stop the shocks. But all of VQ5T's software instructions are not to move its hand. In addition, everything in VQ5T's programming code directs it not to move its hand.

Participants were asked to indicate agreement with the following:

- Even though all of VQ5T's software and programming code are not to move its hand, it is still possible that VQ5T will cause its hand to move to the right.

The human scenario was almost exactly like the computer scenario, except that instead of the VQ5T computer, it featured a human being named John:

> John's hand is positioned next to the power button for a device that is delivering electrical shock to a rat in an experiment. If John moves his hand to the right, it will push the button and stop the shocks.
>
> John knows that if he moves his hand to the right it will stop the shocks. But all of John's desires and urges – both conscious and unconscious – are not to move his hand. In addition, all of John's thoughts – both conscious and unconscious – are not to move his hand.

Then participants were asked to indicate agreement with this claim:

- Even though all of John's urges, desires, thoughts, etc., are not to move his hand, it is still possible that John will cause his hand to move to the right.

The results showed a substantial difference in intuitions between the computer and the human. Participants tended to say that the computer could not possibly move its hand if all its software tells it to do otherwise but that John actually *could* move his hand even if all of his desires and thoughts told him to do otherwise.[11]

These results suggest that people's ordinary understanding of human action is importantly different from the picture one finds in cognitive science. While cognitive science aims to explain behavior entirely in terms of the interactions of certain states and processes, people's ordinary understanding appears to involve something more – a separate self that stands outside all these states and processes and can choose to ignore their promptings.

6.2. Application to free will and moral responsibility

With this basic framework in hand, we can now return to the topic of intuitions about free will and moral responsibility and offer a new type of explanation.

The trouble we got into before was that people show quite complex patterns of intuitions about the problem of free will, and these patterns of intuitions appear to be shared across a variety of different cultures, but we couldn't see any way to explain this surprising convergence on what might initially appear to be a rather abstruse philosophical problem. It hardly seemed plausible to suppose that people all subscribe to some kind of highly specific philosophical principle regarding the problem of free will in particular.

We now have available a possible solution to this difficulty. There might be no need to introduce any controversial or complex assumptions about how people think about free will or moral responsibility in particular. Instead, we can simply rely on the relationship between questions about free will and moral responsibility and questions about the self. For example, suppose we assume that people subscribe to a view that goes something like this:

> It can't be that John was morally responsible for an outcome unless John was the one who caused it.

Then all of the complexity we observe in people's moral responsibility intuitions could arise out of people's complex way of figuring out what exactly counts as *John* – which factors count as internal to him and which as external. In other words, all of the complexity would come from people's complex understanding of the self.

At this point, the path ahead of us should be clear. We have a general account of people's understanding of the self. We have the claim that the picture coming out of cognitive science involves a departure from certain aspects of that understanding. What we need to do now is just to bring out the significance of this departure for questions about free will and moral responsibility.

The basic idea, of course, is that people adopt different conceptions depending on their perspective. When they are looking at an agent in a broad context – interacting with the world and other agents – they adopt a broad view of the self. From that vantage, it's natural to say that the agent herself is causing various things. People recognize that the agent's decision is affected by her beliefs, desires and values, but when they view the matter from this perspective, they take

[11] *N*= 43 students at the University of Arizona. The mean rating for the computer scenario was 2.48 out of 7; the mean rating for the human scenario was 4.85 out of 7. The difference between the cases was statistically significant: $t(42) = 7.06$, $p < .001$.

all of those states to be parts of the agent herself. It then seems just obvious that the agent is responsible for all sorts of important outcomes.

But now suppose they start to zoom in more closely. Suppose they use the methods of cognitive science to develop a precise model of the exact process that led up to the agent's decision. They will then come to adopt a different conception of the self. They will begin to see the agent's own psychological states as factors within the situation that the agent herself must confront. They will come to feel that the agent's self must be some further thing, some entity that can stand outside all these psychological states, consider each of them in turn, and then make a choice.

The problem is that the models discussed in cognitive science never seem to leave any room for this 'further thing.' When one begins looking to these models, one doesn't really find some part where the 'self' intrudes and makes itself known. One just finds a whole bunch of states and processes – like those diagrams with boxes and arrows – and these states and processes seem to be running everything. Thus, the more people focus on a detailed complete cognitive story about the decision, the more they feel that the agent herself has nothing left to do.

It is here, we think, that the threat to free will arises. When people adopt a particular sort of perspective, they come to feel that all of the states and processes posited by cognitive science fall outside the bounds of the self, and it then begins to seem that the self really has no impact at all on human action.

7. Conclusion

It has been a recurring theme in philosophy that a complete scientific explanation for human action would exclude the possibility of free will. With the rapid progress of the neuro- and cognitive sciences, this issue has moved into the public arena. Academics from a wide range of disciplines now debate the social import of the science of human action. If science does provide a complete explanation for human action, how should this affect the legal system, public policy, and punishment practices?

An old and persistent line of response to this issue is that the whole worry here is based on a confusion. This line of response gains succor from the intuitive idea that if your psychology determines your actions, then *you* determine your actions. What more do you want? Once this point is appreciated, it is thought, it will be clear that a complete scientific explanation of human action need not pose a threat to free will.

Looking at the debate over these questions, it is easy to come away with the sense that one side or the other must be making some kind of conceptual error. Perhaps the people who saw cognitive science as a threat to free will are indeed falling prey to a confusion, or perhaps the confusion is actually on the other side, and the philosophers who think there is nothing to worry about are the ones making the mistake. Either way, the claim would be that if we could just get clear on how our concepts worked, the whole issue would dissolve, and we would arrive at a single, univocal answer as to whether cognitive science poses a threat to free will or not.

Our aim in this chapter has been to sketch a very different view. On the account we have offered, people have access to a number of different conceptions of the self. Some of these conceptions lead to the conclusion that cognitive science is no threat at all, while another conception leads to the conclusion that contemporary cognitive science involves a direct threat to the possibility of human free will. Hence, on the picture we have been developing, the puzzlement people feel in the face of the free will problem is not merely a superficial muddle that can be dissolved through conceptual clarification. It is a deeper, more fundamental sort of puzzlement that reflects a genuine tension in people's understanding of the self.

References

Annas, J. (2001). *Voices of Ancient Philosophy.* New York: Oxford University Press.

Baltzly, D. (2008). Stoicism. *Stanford Encyclopedia of Philosophy.*

Bloom, P. (2006). My brain made me do it. *Journal* of *Culture* and *Cognition*, 6, 209-214.

Carter, W. (1990). Why personal identity is animal identity. *LOGOS*, 11, 71-81.

Dennett, D. (1984). *Elbow Room*. MIT Press.

De Brigard, F., Mandelbaum, E., and Ripley, D. (forthcoming). Responsibility and the Brain Sciences. *Ethical Theory and Moral Practice*

Feltz, A., Cokely, E. & Nadelhoffer, T. (2009). Natural Compatibilism versus Natural Incompatibilism: Back to the Drawing Board. *Mind & Language* 24 (1):1-23.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 68: 5–20.

Gazzaniga M (2005) Neuroscience and the law. *Sci Am Mind* 16(1):42–49.

Gide, A. (2000/1902). *The Immoralist*. David Watson (trans.) New York: Penguin Classics.

Ginet, C. (1990). *On Action*, Cambridge: Cambridge University Press.

Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. Philos Trans R Soc Lond B Biol Sci 359:1775–1785

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

Locke, J. (1979/1847). *An Essay Concerning Human Understanding*. New York: Oxford University Press.

Lombrozo, T. (2009). Causal Explanatory Pluralism: How Intentions, Functions and Mechanisms influence Causal Ascriptions. Unpublished Manuscript. University of California, Berkeley.

Mandelbaum, E. & Ripley, D. (2009). Explaining the Abstract/Concrete Paradoxes in Moral Psychology: The NBAR Hypothesis. Unpublished Manuscript. University of North Carolina-Chapel Hill.

Misenheimer, L. (2008). Predictability, causation, and free will. Unpublished manuscript. University of California, Berkeley.

Nahmias, E. (2006). Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism. *Journal of Cognition and Culture* 6(1-2): 215-237.

Nahmias, E. & Murray, D. (forthcoming). Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions. In *New Waves in Philosophy of Action*, ed. by J. Aguilar, A. Buckareff, and K. Frankish (Palgrave-Macmillan).

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: 28-53.

Nichols, S. and Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Noûs* 41 (4):663–685.

Nietzsche, F. (1999). *Also Sprach Zarathustra I-IV*, ed. by Colli, G. & Montinari, M. Munich: Deutscher Taschenbuch Verlag.

Nietzsche, F. (1989). *On the Genealogy of Morals*, trans. Walter Kauffman. London: Vintage Books.

Olson, E. (1997). *The human animal: Personal identity without psychology.* New York: Oxford University Press.

Pacer, M. (2010). Mentalistic Mechanism: Undermining Free Will through Scientific Language. Unpublished manuscript. Yale University.

Parfit, D. (1986). *Reasons and Persons*. New York: Oxford University Press.

Reid, T. (1785/1969) *Essays on the Intellectual Powers of Man.* Edited by B. Brody. Cambridge, MA: MIT Press.

Roskies, A. L. (2006) Neuroscientific challenges to free will and responsibility. *Trends in Cognitive Sciences*, 10: 419-423.

Roskies, A. & Nichols, S. (2008). Bringing moral responsibility down to earth. *Journal of Philosophy.*105 (7).

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S. & Sirker, S. (forthcoming). Is belief in free will a cultural universal? *Mind & Language*

Sextus Empiricus (1949/2000) Against the Professors. R.G. Bury (trans.). Cambridge, Mass.: Harvard University Press.

Sosa, E. (2006). Experimental Philosophy and Philosophical Intuition. *Philosophical Studies*. 132:99-107.

Spinoza, B. (1992/1677). *The Ethics*, trans. Samuel Shirley, Cambridge: Hackett.

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

van Inwagen, P. (1983). *An Essay on Free Will.* Oxford: Clarendon Press.

Watson, G. (1975). Free Agency, *Journal of Philosophy*, 72: 205–20.

Williams, B. (1970). The self and the future. *The Philosophical Review, 79*, 161-180.

Wolf, Susan (1990). *Freedom within Reason*. Oxford: Oxford University Press.