FREE WILL AND THE SCIENTIFIC VISION[1]

Joshua Knobe
Yale University

[Forthcoming in Edouard Machery and Elizabeth O'Neill (eds.),
Current Controversies in Experimental Philosophy. Routledge.]

Take a few courses in cognitive science, and you are likely to come across a certain kind of metaphor for the workings of the human mind. The metaphor goes like this:

> Consider a piece of computer software. The software consists of a collection of states and processes. We can predict what the software will do by looking at the complex ways in which these states and processes interact.
>
> The human mind works in more or less the same way. It too is just a complex collection of states and processes, and we can predict what a human being will do by thinking about the complex ways in which these states and processes interact.
>
> One can imagine someone saying: 'I see all these states and processes, but aren't you forgetting something further – namely, the person herself?' This question, however, is a foolish one. It is no more helpful than it would be to say: 'I see all these states and processes, but where is the software itself?' The software just *is* a collection of states and processes, and the human mind is the same sort of thing.

This metaphor does a good job of capturing the basic approach one finds throughout the sciences of the mind. We might say that it captures the *scientific vision* of how the human mind works.

But one could also imagine another, very different metaphor for the workings of the mind. Perhaps something like this:

> Consider a royal court. The advisors and ministers each have an opportunity to advocate for a particular course of action. But it is not as though the advisors and ministers themselves make the final decision. Instead, there is another person in the court – the king or queen – who listens to all of the arguments, thinks them over, and then decides.
>
> The mind works in more or less the same way. Your mind might include various states and processes, but it would be a mistake to suggest that you yourself are just a collection of states and processes. On the contrary, you are a further thing – like the king or queen in the court – who can attend to the states and processes within your mind and then freely make a choice.

[1] The theory presented here was developed in close collaboration with Shaun Nichols (as should be clear from sections 1 and 2), and a number of aspects of it were directly inspired by the research of Eddy Nahmias and Dylan Murray (see sections 3 and 4). I am deeply grateful to all three of these philosophers, both for their published research and for numerous invaluable conversations.

> When you do end up making a free choice, we might say that you made this choice 'on the basis of' some of your psychological states. But the connection here is always indirect. It is not as though your psychological states actually *cause* your behavior; you just freely decide what to do, and sometimes you end up deciding to act in a way that accords with them.

On this latter metaphor, the self is something that transcends all of the states and processes within the mind. Indeed, it is something that transcends the whole causal order. We might therefore refer to this second view as the *transcendence vision*.

A question now arises as to which of these two visions best captures people's ordinary understanding of human action. Thus, suppose that in the course of a perfectly ordinary conversation, someone utters the sentence:

'John went to New York because he wanted to visit his sister.'

Presumably, the person uttering this sentence thinks of John as choosing to go to New York with *free will*. That is, the person assumes that John freely decided to perform this action and was not in any way compelled. Nonetheless, the sentence quite clearly states that John's action can be explained in terms of his psychological states. Specifically, the sentence says that John performed this action (going to New York) because he had a particular desire (to visit his sister). So it seems that we are faced with a problem. How exactly do people ordinarily understand the role of psychological states in cases of free action? Can we capture people's ordinary understanding in terms of something like the scientific vision, or do we need to invoke something more along the lines of the transcendence vision?

In existing work on people's understanding of mind and action, it is common for researchers to ignore the whole issue of free will and to assume that people's ordinary understanding conforms, at least in broad outlines, to the scientific vision. Thus, it has been said that people's ordinary understanding works something like a scientific theory, that people try to understand human action by looking for its causes and, in particular, that they think of human action as caused by psychological states. On this sort of view, the contemporary scientific study of human cognition isn't really too much of a departure from people's ordinary way of making sense of the mind. It is just a more precise, systematic way of doing the very same thing that people do all the time. (For discussion, see Bloom 2006; Churchland 1981; Gopnik & Wellman 1992; Lewis 1972; Nichols 2006.)

I will be arguing that this view is mistaken. I suggest that people's ordinary way of making sense of the mind conforms more to the transcendence vision. Hence, the approach that we find in contemporary cognitive science is not just a more precise or systematic way of doing the same thing we do all the time. On the contrary, the basic vision at the heart of that approach is actually incompatible with people's ordinary way of understanding human freedom.

The evidence for this claim comes primarily from empirical work on the nature of people's ordinary understanding. Accordingly, we will be looking at a series of different empirical studies. Each study might be somewhat inconclusive on its own, but together, they form a powerful and surprisingly coherent package.

1. Let us begin with the most straightforward and obvious method for addressing these questions. If we want to know how people ordinarily understand human behavior, one approach would be to start by just asking them directly.

A few years ago, Shaun Nichols and I conducted an investigation using precisely this approach (Nichols & Knobe 2007). Participants were told to imagine two possible universes. First, they were introduced to a universe in which every event was caused by some prior event:

> 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.

Then they were told about a second universe that was similar to the first in many ways but that differed in one crucial respect.

> 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.

So in the first universe, every event is caused by some prior event, while in the second universe, human decisions are not caused by any prior event at all; they are just freely chosen.

After reading about these two possible universes, participants were asked:

> Which of these universes do you think is most like ours? (circle one)
>
> Universe A Universe B

Let us now introduce a helpful abbreviation. Instead of saying that an event is 'completely caused by whatever happened before it,' we will say that it is *causally determined*. We can then state the difference between the two universes as follows: In Universe A, everything is causally determined, whereas in Universe B, human actions are not causally determined. Which type of universe do people think is most like ours?

The overwhelming answer is: Universe B. In our original studies, this answer was chosen by over 90% of participants. In other words, when people were asked directly, they tended to say that our universe was like a universe in which almost all events were causally determined but human actions were not. This result suggests that people's explicit answers fit more closely with the transcendence vision than with the scientific vision.

Most researchers working in this area would agree that this type of evidence is far from conclusive. They would say: 'Our aim is to investigate the tacit mechanisms that people use all the time to understand human action. The best way to figure out how these mechanisms work is to look at various aspects of people's thinking (their moral judgments, their explanations, etc.) and try to use these aspects of people's thinking as clues to the nature of their most basic understanding. The technique you've used here is far less reliable. If you insist on just directly asking people abstract questions about the nature of action, there is no guarantee that you will be tapping into these mechanisms in any way. The answers people give might simply reflect the explicit theories they have picked up in various philosophical conversations. (Maybe they were exposed to ideas from Christian

theology in their Sunday school classes, and they are simply parroting back something they've learned there.)'

This is a powerful objection, and it deserves to be taken very seriously. To address it, we teamed up with Hagop Sarkissian and tried running a cross-cultural study. Participants were recruited from the United States, Hong Kong, India and Colombia. All of these participants were then given precisely the same question described above (Sarkissian, Chatterjee, De Brigard, Knobe, Nichols & Sirker 2010). The results showed no significant differences between cultures. In all four cultures, the majority of participants said that our own universe is most similar to the one in which human action is not causally determined.

This result comes as something of a surprise. Here we have people from radically different cultures, with quite different historical and religious traditions, and yet they all seem to be converging on the same answer to this highly abstract question. How are we to explain this convergence? It hardly seems plausible to suggest that all of these people have been taking classes in which they are explicitly taught the same philosophical theory. Presumably, we need to provide some other type of explanation.

If we assume that people's ordinary understanding of action follows the transcendence vision, this task becomes quite simple. The explanation is that the experiment is accurately tapping into people's ordinary understanding. (This understanding tells people that human action is not causally determined, and they answer the explicit theoretical questions accordingly.) By contrast, if we assume that people's ordinary understanding follows the scientific vision, the matter becomes considerably more complex. We would need to argue that people's tacit understanding is telling them that human action actually *is* causally determined but that there is some further factor – a factor that is equally present in all four of these cultures – which then obscures this tacit understanding and leads people to explicitly state that human action is *not* causally determined. It would require some ingenuity to develop an explanation along these lines, but perhaps future research will lead to the development of a theory that can successfully pull off this trick.

2. A question now arises about how people make judgments of moral responsibility. Do people think that an agent can be morally responsible for her behavior if this behavior is causally determined? Or do people think that a person can only be truly responsible if her behavior is freely chosen by a transcendent self?

A number of studies have examined this question, and these studies have yielded a surprising result. Suppose we tell participants about an agent whose behavior is causally determined. Now suppose we tell them that this agent performs some immoral behavior. We might tell them that the agent has committed rape, or that he has robbed a bank, or that he has murdered an innocent person. Participants who have been given cases of this form tend to arrive at a striking sort of moral judgment. Even though they have been informed in no uncertain terms that the agent in the case is causally determined, they tend to say that the agent is fully morally responsible!

This result was first uncovered in an influential paper by Nahmias, Morris, Nadelhoffer and Turner (2006), and it has since been replicated and extended by numerous other researchers (De Brigard, Mandelbaum & Ripley 2009; Feltz & Cokely 2009; Nahmias, Coates & Kvaran 2007; Nichols & Knobe 2007). At this point, the basic finding has been established beyond all reasonable doubt.

In one of the most impressive experimental demonstrations of this phenomenon (De Brigard, Mandelbaum & Ripley 2009; Mandelbaum & Ripley forthcoming), participants were told to imagine a person named Dennis. They were informed that Dennis had a neurological disorder, that the bad things he did were completely caused by this neurological disorder, and that if other people were to have the same disorder, they would do the very same bad things. After receiving this information, participants were told that Dennis had raped a number of women, and they were asked whether he was morally responsible for his actions. Surprisingly, participants tended to say, even in this very extreme case, that Dennis actually *was* responsible.

Now, one possible reaction to these results would be to say that people's understanding of moral responsibility is more or less independent of questions about causal determinism. But Nichols and I thought that there might be more to the story. Our hunch was that people's intuitions in these cases might be the product of two different psychological processes that were pulling them in opposite directions. More specifically, we thought that (a) when people are engaged in abstract theoretical reasoning, they use a conception according to which agents cannot be morally responsible for behavior that is causally determined, but then (b) when they hear about some specific concrete act of wrongdoing (rape, robbery, murder, etc.), a further process comes into play that leads them to say that the agent actually is responsible for his misdeeds.

To test this hypothesis, we conducted an additional study (Nichols & Knobe 2007). All participants received the description of the causally deterministic Universe A and were asked whether people within this universe could be morally responsible. But the study included an additional wrinkle: participants were randomly assigned to be asked this question in one of two possible ways. Participants in the 'concrete' condition received the question:

> 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

In this condition, most participants (72%) gave the answer 'Yes,' indicating that a causally determined agent could still be morally responsible. This first result simply replicates earlier findings.

Then, in the 'abstract' condition, participants received the question that did not mention any actual concrete misdeeds:

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

In this latter condition, participants gave exactly the opposite pattern of responses, with the vast majority (86%) choosing the answer 'No.'

This difference between concrete and abstract judgments is a puzzling one, and the attempt to understand it has been one of the major preoccupations of experimental philosophy work on free will. There have been studies examining the phenomenon more systematically using a variety of different descriptions of determinism (Nahmias, Coates & Kvaran 2007), studies manipulating people's level of abstract thinking by asking them to think about either close or distant times (Weigel 2011), even studies that go after these questions by manipulating the *font* in which the stimuli are written (Gonnerman, Reuter & Weinberg 2012). A number of competing theoretical models have been proposed (Cova, Bertoux, Bourgeois-Gironde & Dubois 2012; Nahmias & Murray 2010; Nichols & Knobe 2007), but at this point, no clear consensus has emerged. Perhaps future work will bring more clarity to these issues.[2]

For present purposes, however, we can focus on a slightly different question. Instead of asking why people are more inclined to regard the agent as responsible in concrete cases, we can ask why people do not regard the agent as obviously responsible in all of these cases. There is clearly something drawing people to the view that an agent can't be morally responsible for behaviors that are causally determined. But what exactly is drawing people in this direction? Why do they see causal determinism as any problem at all for moral responsibility?

At this point, one might offer a number of different hypotheses, but it is a striking fact that people's worry about causal determinism can be very easily explained if we simply assume that people accept the transcendence vision. It then becomes unnecessary to make any complex further assumptions about the way people think about moral responsibility in particular. All one needs is the straightforward principle:

> People will be reluctant to hold an agent responsible for a behavior if they believe that this behavior was not even produced by the agent in question.

---

[2] In earlier work, Shaun Nichols and I suggested that this effect might be arising because people have an emotional reaction to the concrete case and this reaction biases their responses (Nichols & Knobe, 2007). Over the past few years, this hypothesis has been put to the test in a variety of different experiments, and at this point, I have to say that things are not looking good. First, a series of studies looked at cases that were entirely concrete but which differed in the extent to which they would be expected to provoke emotional reactions. A recent meta-analysis of 29 such studies shows that people are indeed more inclined to ascribe moral responsibility in a deterministic universe when faced with a high-affect case than with a low-affect case but that this effect is quite small (d = .18) – not nearly large enough to explain the powerful impact of the original abstract/concrete manipulation (Feltz & Cova, 2012). Second, a recent study looked at the responses given to abstract and concrete cases among participants with frontotemporal dementia. Though these participants show a deficit in their capacity for emotional response – and would therefore be expected to differ from neurotypical participants if the effect was driven by emotion – they ended up giving exactly the same pattern of responses seen in earlier studies (Cova, Bertoux, Bourgeois-Gironde & Dubois, 2012).

In light of these results, I now suspect that the abstract/concrete effect is not, in fact due to emotional responses. Perhaps it can be explained instead by the very theory proposed here: People have a very strong tendency to think of human decision-making in terms of transcendence. Thus, no matter how much one emphasizes determinism, if the case is presented with sufficient concreteness, participants immediately apply their default (transcendence-based) framework.

The key result then follows almost immediately. If the transcendence vision is correct, then all behaviors that are causally determined – even if they are determined by the agent's own psychological states – will not truly have been produced by the agent herself.

We now arrive at a provisional conclusion. Experimental results show that people see causal determinism as a threat to moral responsibility, and any correct theory here will have to explain why people have this intuition. If we start out with the view that people accept the scientific vision, we might be able to offer an explanation by introducing certain further assumptions. But the situation becomes very different if we start out with the view that people accept the transcendence vision. It then becomes possible to explain the results without introducing any further controversial assumptions. Everything simply follows from people's understanding of what would be required for an agent even to have produced the action at all.

3. In an ingenious series of studies, Nahmias and Murray (2010) looked more directly at the ways in which people's understanding of psychological states impacts their intuitions about free will. Once again, participants were told about the causally deterministic Universe A, but this time, they were asked whether they agreed or disagreed with statements of the form:

- In Universe A, what a person believes has no effect on what they end up being caused to do.
- In Universe A, what a person wants has no effect on what they end up being caused to do.

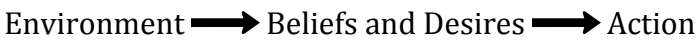
Surprisingly, participants tend to *agree* with these statements. In other words, when participants are informed that an agent's actions are causally determined, they tend to infer that the agent's actions do not depend in any way on her beliefs and desires. This finding constitutes a genuine breakthrough within research in this area, and it is worth taking the time to think in detail about what it might be telling us.

To get a better sense for the broader implications of the Nahmias-Murray findings, it might be helpful to introduce an analogy. Suppose we are looking at a house that has been destroyed, and I tell you: 'This house burned down in a fire.' Now suppose that a little while later you receive one further piece of information: 'The destruction of the house was completely caused by an event that occurred three days ago.' Presumably, you would not conclude that the fire had no effect at all on what happened to the house. Instead, you would probably assume that the fire was precisely the event that occurred three days ago and completely caused the destruction. (You might then infer that if the fire had never occurred, the house would still be in fine shape today.) What the Nahmias-Murray results show is that people do not apply this same kind of reasoning when it comes to the relationship between human action and mental states. On the contrary, when people are told that an agent's actions are completely caused by prior events, they conclude that the agent's beliefs and desires could not possibly be having any effect on what she ends up doing. This result seems to suggest that people are conceptualizing the relationship between an agent's beliefs and desires and her actions in a way that is radically different from the way they would normally conceptualize the relationship between a fire and the destruction of a house. Our aim now is to get a better sense for the nature of that difference.

Let us begin by stating the obvious. Clearly, people often explain why an agent acted in the way she did by referring to her beliefs and desires. Thus, if we pick out an agent's action and ask 'Why did she do that?' we might receive an answer like: 'Because she believed that it was the right thing to do.' Then, continuing the chain of explanation back a step, people often explain an agent's beliefs and desires by tracing them to facts about her external environment. So if we ask the question 'Why did she believe it was the right thing to do?' we may receive the answer: 'Because her parents always told her that it was right to behave in that way.'

Yet, though it is perfectly clear that people offer explanations of this form, it has proven remarkably difficult to say precisely what these explanations *mean*. In the mid-twentieth century, a great deal of philosophical work went into trying to understand the sentences people use, in ordinary language, to explain an agent's actions in terms of her beliefs and desires. Philosophers developed a variety of opposing theories (Davidson 1963; Hart & Honoré 1959; Ryle 1949; Wittgenstein 1953), but the issue was never fully resolved. In my view, the question is just as mysterious today as it was when it was first posed.

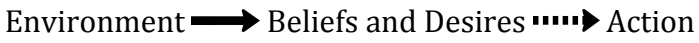
One obvious answer would be that the relationship between an agent's beliefs and desires and her actions is best understood as a straightforward case of *causation* (e.g., Davidson 1963). On this view, what we have is a causal chain: the agent's environment causes her to have certain beliefs and desires, which in turn cause her to perform a particular action.



Many researchers working on these issues will immediately feel that this picture is clearly the correct one. In fact, some researchers may find themselves hard-pressed even to imagine an alternative.

But if we start out with the assumption that people accept a picture like this one, we soon run into a major difficulty. According to the picture, most people's actions are causally determined by their beliefs and desires. It should therefore be blazingly obvious that even if an agent's actions are causally determined, her beliefs and desires can have an effect on what she does. (After all, the assumption is that people are thinking that, in a typical case, it is precisely the agent's beliefs and desires that causally determine her actions.) But now we face the problem. For what the Nahmias-Murray results show is that people *don't* endorse this conclusion. Instead, people tend to say that if an agent's actions are causally determined, her beliefs and desires cannot have any effect at all on what she does. Why might people be responding in this way?

One natural answer would be that people reject the whole picture we have been sketching. Instead, they might accept something along the lines of the transcendence vision. On this latter picture, the relationship between an agent's psychological states and her actions is not a simple causal chain. It is something more complex:



Here the dotted line signifies a relationship that involves not straightforward causal explanation but rather what philosophers call 'reason explanation.' In other words, on this view, when people explain an action in terms of a belief, they are not saying that the action

was *caused* by the belief. Rather, they are saying that the action was *chosen for a reason*. (For example, they might be saying that the agent's reason for choosing the action was her belief that it was the right thing to do.)

It would be difficult to say exactly what it means for an agent to do something 'for a reason' in the relevant sense. Different theorists have developed quite different accounts (Anscombe 1957; Hart & Honoré 1959; Wittgenstein 1958:14-15), and the debate continues up until the present day (Aguilar & Buckareff 2010; Knobe 2007; Malle 2004; Mele 2010). For present purposes, however, we do not need to resolve that controversy. The key idea is just that, on the transcendence vision, an agent can do something for a reason even when the resulting action was freely chosen and not caused by anything at all.

If we start out with this sort of framework, it becomes easy to see why people might respond as they do in the Nahmias-Murray experiments. People think that when an agent acts on the basis of reasons, her behaviors are not causally determined. Then they are told about a universe that differs from our own in that everything in it actually is causally determined. They therefore infer that agents in this universe do not do things for reasons. It is only a small step from this inference to the conclusion that the beliefs and desires of these agents have no effect on what they end up doing.

By contrast, suppose we start out with the assumption that people's ordinary understanding is well captured by the scientific vision. Now our starting assumption is that people think of ordinary human action as causally determined by psychological states. We then learn a new fact. When people are told about a universe in which all human actions are causally determined, they conclude that the actions of human beings in this universe do not depend in any way on their psychological states. How on earth are we to explain this fact? It is certainly possible that someone will come up with a viable solution here, but the problem is not looking like an easy one.

4. At the heart of the transcendence vision is the idea that human actions are radically different from other sorts of events. We might explain the movements of a billiard ball by saying that its movements were caused by prior events, but the explanation of a human action would have to be entirely different. Human actions are not *caused* by prior events; they are *chosen* on the basis of reasons.

This point comes out especially clearly when we consider events that might in some ways seem similar to human actions. Take the case of computers. One might think that the explanation of a computer's output ought to resemble, at least in certain minimal respects, the explanation of a human action. (Computers contain internal representations, and their outputs can be explained in terms of those representations.) Yet, even on the transcendence vision, the output of a computer will be best understood as the product of a perfectly straightforward causal chain:

Environment ⟶ Program ⟶ Output

No matter what you think about the vexed questions surrounding human action, there is little temptation to suppose that anything equally mysterious is occurring in the case of computers. A computer does not proceed by considering its own program and then freely choosing which output to display. Rather, the program simply *causes* the computer to generate a particular output.

Indeed, one should be able to see this sort of straightforward causal chain even in certain kinds of cases in which a human being's bodily movements are explained in terms of her psychological states. Suppose that an agent is watching a scary movie and makes an involuntary grimace. The process might then go like this:

Environment ⟶ Emotions ⟶ Facial Expressions

The transcendence vision suggests that cases like this one are deeply different from cases of voluntary action. Voluntary actions might be seen as freely chosen on the basis of reasons, but clearly, no such thing is taking place in a case like this one. It is not as though the agent freely chooses to grimace and her reason for making that choice is that she is afraid. Rather, the fear directly *causes* the grimace. Here again, nothing more complex or mysterious is required.

In short, the transcendence vision leaves us with the idea that there is a fundamental difference between different kinds of explanation. It says that people's way of explaining free human action should be deeply different from their way of explaining a computer's behavior or an involuntary grimace. But in that case, it seems that we immediately arrive at a new testable prediction. If people's way of understanding free human action is radically different from their way of understanding other sorts of events, and if the Nahmias-Murray experiments do indeed give us a way of tapping into people's understanding, then we should be able to use a modified Nahmias-Murray experiment to show a difference between judgments about free human action and judgments about other sorts of events.

To put this prediction to the test, I conducted a quick follow-up study. All participants received the story about Universe A. Participants in the 'human' condition then received the following question:

> Imagine that Universe A includes various people who have beliefs and values. Now please tell us whether you agree or disagree with the following statement:
>
> - In Universe A, people's beliefs and values have no effect on what they end up being caused to do.

Meanwhile, participants in the 'computer' condition received the question:

> Imagine that Universe A includes various computers that use programs and data. Now please tell us whether you agree or disagree with the following statement:
>
> - In Universe A, the computers' programs and data have no effect on what they end up being caused to do.

Participants tended to agree with the statement in the human condition but to disagree with the statement in the computer condition. [3]

[3] Forty-one people were recruited through Amazon's Mechanical Turk. Ratings were on a scale from 1 ('disagree') to 7 ('agree'). Agreement was higher in the human condition ($M = 5.4$) than in the computer condition ($M = 3.6$), $t(39) = 2.4$, $p < .05$.

Notice the striking pattern of intuitions people are showing in this case. They are saying that if everything in the universe is causally determined, then a computer's data can still have an effect on its output but a human being's beliefs cannot have any effect on her behavior. This is exactly the result one would predict if one started out with the assumption that people are adopting the transcendence vision, and it is hard to see how one would explain it on any other hypothesis.

To further get at the nature of the effect here, I conducted a second study using an even more closely controlled pair of cases. This time, participants in the 'reason' condition received the following question:

> Imagine that the people in Universe A perform various actions. Now please tell us whether you agree or disagree with the following statement:
>
> - In Universe A, people's beliefs and values have no effect on what actions they end up performing.

Other participants were assigned to the 'non-reason' condition:

> Imagine that the people in Universe A make various facial expressions. Now please tell us whether you agree or disagree with the following statement:
>
> - In Universe A, people's emotions have no effect on what facial expressions they end up making.

Here again, there was a significant difference between conditions. People tended to agree in the reason condition but not in the non-reason condition.[4] What we have here, then, is an even more tightly controlled minimal pair. If the universe is completely deterministic, people think that an agent's emotions can still impact her facial expressions but that an agent's beliefs cannot impact her actions. Once again, these are exactly the results one would expect if one started out with the view that people accept the transcendence vision.

Of course, if we start out with the idea that people accept the scientific vision, we might be able to develop an alternative explanation for these findings, but this is beginning to look like a losing battle. To hold onto the hypothesis that people accept the scientific vision, we would need to develop an alternative explanation for the findings about people's explicit judgments about the nature of our universe, *and* for the findings about moral responsibility judgments, *and* for the Nahmias-Murray findings, *and* for these new findings about the way people's judgments about emotions differ from their judgments about beliefs. Why would we be at all drawn to pursue a research program along these lines?

5. Clearly, the claim that people accept something like the scientific vision is not just a single isolated hypothesis. It is a natural part of a larger picture of how the human mind works, and one might think that the only way to really do justice to this claim is to understand it in the context of the larger picture in which it is embedded.

The larger picture says that people's basic way of making sense of the world is something more or less like a scientific theory. This picture has been developed in rich

[4] Forty people were recruited through Amazon's Mechanical Turk. Ratings were on a scale from 1 ('disagree') to 7 ('agree'). Agreement was higher in the reason condition (M = 5.7) than in the non-reason condition (M = 3.5), $t(38) = 3.3$, $p < .005$.

theoretical detail, and it has been applied in research on everything from people's causal judgments (e.g., Gopnik, Glymour, Sobel, Schulz, Kushnir & Danks 2004) to their understanding of psychological states (e.g., Churchland 1981). One can then apply this general picture quite straightforwardly to the question under discussion here. From a more scientific perspective, the transcendence vision looks a bit spooky, perhaps even conceptually incoherent. So if one starts with the idea that people's way of making sense of the world is a broadly scientific one, it may begin to seem just obvious that people have to accept something along the lines of the scientific vision.

In my view, this type of argument is a very powerful one. If we find a general picture that ends up generating accurate predictions in one domain after another, there is certainly strong reason to suspect that it will continue to prove accurate when we switch over to a new domain. Indeed, even if we run into some difficulties in this new domain, it might be reasonable to try dismissing those difficulties and sticking with the general picture. One might say: 'This general picture turned out to be correct in so many other cases. We might be running into some troubles right at the moment, but if we stick with our general research program, we will surely be able to work them out in the end.' Such a response could, in the right circumstances, be exactly the right one.

But, of course, the argument cuts both ways. One of the major results of existing work in experimental philosophy is that when one looks in detail at people's ordinary intuitions, one finds that these intuitions look very different from anything one would expect to find in the sciences. When one looks at intuitions about happiness or knowledge, one finds that these intuitions are shaped by moral considerations (e.g., Beebe & Buckwalter 2010; Phillips, Nyholm & Liao forthcoming). When one looks at intuitions about mental states, one finds that they do not conform to the kind of functionalist approach found in cognitive science but instead take our embodiment into account (Huebner 2010; Knobe & Prinz 2008; cf. Sytsma & Machery 2009). When one looks at intuitions about causation, one finds that these intuitions follow patterns that are deeply different from those involved in scientific causal modeling (Alicke, Rose & Bloom 2011; Hitchcock & Knobe 2011; Sytsma, Livengood & Rose forthcoming). Similar results have been found in numerous other domains (for a review, see Knobe 2010).

So perhaps we can now turn the argument around and run it in the other direction. We have a general research program of investigating the ways in which people's ordinary intuitive understanding is deeply different from the kind of understanding one finds in the sciences, and this research program has generated accurate predictions in numerous other domains. We now have good reason to suspect that this general program will prove helpful in the present case as well.

Conclusion

This chapter has been concerned with questions about how people ordinarily understand free human action. Consider again the sentence: 'John went to New York because he wanted to visit his sister.' This sentence seems to suggest that John freely chose to perform an action on the basis of certain reasons. How exactly do people ordinarily make sense of this notion?

One plausible view would be that we can capture the ordinary understanding of free action using more or less the approach found in contemporary cognitive science. Perhaps our sentence means that John had a desire to visit his sister and that this desire figured in a

complex cognitive process that eventually caused him to go to New York. To the extent that we want to capture the notion that he performed this action freely, we might then invoke various other cognitive scientific concepts (self-regulation, cognitive control, etc.).

I have argued that this approach is misguided. Our ordinary way of making sense of free action is deeply different from anything that appears in cognitive science and, indeed, from anything in the sciences more generally. So as long as we are working within a broadly scientific framework, we will never be talking about the very thing that people are getting at with their ordinary notion of free will.

Of course, in making this claim about the ordinary notion, I do not mean to be ruling out the possibility of a scientific theory of free will. We might well find in the end that we have reasons of one kind or another to conclude that certain cognitive phenomena actually are sufficient for free action. But if we do go down this route, we should be clear about what we are doing. We will not be simply working out the implications of the ordinary understanding of free action. Instead, we will be abandoning this ordinary notion and replacing it with a very different one.

References

Aguilar, J. H. & Buckareff, A. A. (eds.) (2010). *Causing human actions: New perspectives on the causal theory of action*. Cambridge: The MIT Press.

Alicke, M., Rose, D. & Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy* 108: 670-696.

Anscombe, G.E.M. (1957). *Intention*, Oxford: Basil Blackwell.

Beebe J. & Buckwalter W. (2010). The epistemic side-effect effect. *Mind & Language* 25:474–9.

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

Churchland, P. (1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78: 67-90.

Cova, F., Bertoux, M., Bourgeois-Gironde, S. & Dubois, B. (2012). Judgments about moral responsibility and determinism in patients with behavioural variant of frontotemporal dementia: Still compatibilists. *Consciousness and Cognition* 21: 851-864.

Davidson, D. (1963). Actions, Reasons and Causes. *Journal of Philosophy*. 60: 685–700.

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

Feltz, A. & Cokely, E. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition* 18: 342-350.

Feltz, A. & Cova, F. (2012). When and how affective reactions impact judgments about free will and determinism: A meta-analysis. Unpublished manuscript. Schreiner University.

Gonnerman, C., Reuter, S. & Weinberg, J. (2012). More oversensitive intuitions: Print fonts and could choose otherwise. Unpublished manuscript. Indiana University.

Gopnik, A. & Wellman, H. (1992). Why the child's theory of mind really *is* a theory. *Mind & Language* 7: 145-171.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111: 1-31.

Hart, H. L. A. & Honoré, T. (1959). *Causation in the Law*. Oxford: Clarendon.

Hitchcock C. & Knobe J. (2011). Cause and norm. *Journal of Philosophy* 106:587—612.

Huebner B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone *care* about functional zombies? *Phenomenology and the Cognitive Sciences* 9: 133—55.

Knobe, J. (2007). Reason explanation in folk psychology. *Midwest Studies in Philosophy*, 31: 90-107.

Knobe J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences* 33:315—29.

Knobe J. & Prinz J. (2008). Intuitions about consciousness: experimental studies. *Phenomenology and the Cognitive Sciences* 7: 67—83.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50: 249-258.

Malle, B. F. (2004). *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, MA: MIT Press.
Mandelbaum, E. & Ripley, D. (forthcoming). Explaining the abstract/concrete paradoxes in moral psychology: The NBAR hypothesis. *Review of Philosophy and Psychology*.
Mele, A. (2010). Teleological explanations of actions: Anticausalism versus causalism. In Aguilar, J. H. & Buckareff, A. A. (eds.) 2010. Causing human actions: New perspectives on the causal theory of action. Cambridge: The MIT Press, 183–198.
Nahmias, E., Coates, D., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy* 31: 214—42.
Nahmias E., Morris S., Nadelhoffer T., & Turner J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73: 28—53.
Nahmias E. & Murray D. (2010). Experimental philosophy on free will: an error theory for incompatibilist intuitions. In *New Waves in Philosophy of Action*, ed. J Aguilar, A Buckareff, K Frankish, pp. 189 – 216. Hampshire, UK: Palgrave-Macmillan.
Nichols, S. (2006). Folk intuitions about free will. *Journal of Cognition and Culture* 6: 57-86.
Nichols S. & Knobe J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous* 43: 663 – 85.
Phillips, J., Nyholm, S. & Liao, S. (forthcoming). The good in happiness. *Oxford Studies in Experimental Philosophy*.
Roxborough, C. & Cumby, J. (2009) Folk psychological concepts: Causation. *Philosophical Psychology* 22: 205-13.
Ryle, G. (1949). *The concept of mind*. London: Hutchinson.
Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S. & Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language* 25 (3):346-358.
Sytsma J. & Machery, E. (2009). Two conceptions of subjective experience. *Philosophical Studies* 151: 299—327.
Sytsma, J., Livengood, J. & Rose, D. (forthcoming). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Science*.
Uttich, K. & Lombrozo, T. (2010). Norms inform mental state ascriptions: a rational explanation for the side-effect effect. *Cognition*, 116: 87-100.
Weigel C. (2011). Distance, anger, freedom: an account of the role of abstraction in compatibilist and incompatibilist intuition. *Philosophical Psychology* 24:803-823.
Wittgenstein, L. (1958). *The blue and brown books*. New York: Harper & Row.
Wittgenstein, L. (1953) *Philosophical investigations*, Translated by G. E. M. Anscombe. Oxford: Basil Blackwell.