**A Naturalistic Vision of Free Will**

Eddy Nahmias and Morgan Thompson[[1]](#footnote-1)

Georgia State University

[Draft of a chapter for Elizabeth O’Neill and Edouard Machery (eds.),

*Current Controversies in Experimental Philosophy*. Routledge.]

Consider a future brain imaging technology that would allow neuroscientists to see all of your brain activity, and with this information they could know what you were going to think, decide, and do before you were even aware of making your decisions. While wearing this brain scanner (imagine it is a lightweight cap), the scientists can see, for instance, how you will vote in an election before you are aware of making your final decision. They make these predictions with 100% accuracy. Ask yourself: Is such technology possible (even if not in our lifetime)? If it *were* possible, would it show that each of our decisions is caused by particular brain states? And would such technology show that we lack free will and moral responsibility?

A scenario very much like this one is used by Sam Harris in his short book *Free Will* in order to “expose this feeling [of free will] for what it is: an *illusion*” and to help his readers understand that “the laws of nature [are] incompatible with free will” (2012: pp. 10-11).[[2]](#footnote-2) This scenario might also suggest the “scientific vision” that Joshua Knobe argues, in the previous chapter, is incompatible with people’s understanding of free will. It certainly seems to conflict with the “transcendence vision” as he defines it. If there were a “transcendent self,” something that “transcends all the states and processes in the mind” and indeed “the whole causal order” (1), then how could information about brain activity provide *complete* information, and perfect predictions, about what a person will think and do, before the person (the “transcendent self”) is even aware of what he or she will decide? What role could such a transcendent self play in action?

If Harris and Knobe are accurately gauging ordinary people’s views, then they should predict that most people will say that this brain imaging technology could not be developed—that it would inevitably miss crucial information about what the transcendent self is thinking and deciding—and that if the technology did exist, it would rule out free will and responsibility. However, these predictions fail.

We presented people with a detailed explanation of this technology (see scenario below). Most people responded that it could be developed, and roughly four out of five said that if it existed it would *not* rule out free will or responsibility. We will argue that these results, as well as other experimental philosophy results, including those discussed by Knobe, are best explained by a different theory of how most people understand the self and free will.

Here’s the plan for our chapter: First, we will describe a “naturalistic vision” of free will that represents a middle ground between the extremes Knobe describes as the “scientific vision” and the “transcendence vision.” We will briefly explain why the naturalistic vision offers a plausible metaphysical theory that is consistent with what science tells us (so far) about the mind and that is also consistent with most people’s beliefs about the self and mind. We will then present people’s responses to our brain-scan scenario and our interpretations of these results. Next, we will offer our interpretation of other results in experimental philosophy and introduce our “causal competition principle” to explain why people interpret some scenarios as threatening free will and responsibility. This principle, we will conclude, also suggests that the plausibility of the naturalistic vision to laypersons will depend on whether a viable future scientific theory of mind fills out the naturalistic vision in a way that does *not* entail causal competition between conscious mental activity and brain activity.

**1. The Naturalistic Vision as a Theory-Lite Compromise**

Naturalism is a fuzzy term. That serves our purposes well, since we think that ordinary people have a fuzzy understanding of the nature of mind, action, and free will. Their theoretical commitments are relatively non-committal and revisable. In contrast, both Knobe’s “scientific vision” and his “transcendence vision” have more robust theoretical commitments, so we think a compromise between them will better capture most ordinary people’s views.

Accepting naturalism commits one to the ontological thesis that whatever exists is composed of things that could be studied by the natural sciences and that conform to the laws of nature. As such, it does conflict with Knobe’s dualistic transcendence vision, which posits an agent who transcends the causal order and makes choices while “not caused by anything at all” (8).

Naturalism is compatible with various forms of physicalism in philosophy of mind, including both non-reductive and reductive varieties (Stoljar 2009). However, naturalism does not commit one to a reductionistic ontological thesis that says the only things that *really* exist are whatever entities physics determines compose everything (such as subatomic particles or strings), nor to a reductionistic epistemological thesis that says the best explanations will be those offered by lower-level sciences (e.g., physics or neuroscience). As such, it is consistent with Knobe’s general description of the “scientific vision,” though it is not constrained by the reductionistic computer science metaphor he uses to introduce it.

Indeed, the naturalistic vision is consistent with a more complicated but less theoretically committed folk psychology than Knobe’s scientific vision suggests. It is consistent with agents’ having capacities for self-reflection, for consideration of reasons for various courses of action and caring about some more than others, and for efforts to control one’s decisions and actions in light of deliberation. The naturalistic vision is consistent with contemporary compatibilist accounts of free will as involving, for instance, identification (Frankfurt 1992), a “deep self” (Wolf 1990), and reasons-responsiveness (Fischer and Ravizza 1992). Indeed, it is consistent with the sort of deliberative processes Knobe initially suggests in his metaphor for transcendence, in which the king or queen “listens to all the arguments, thinks them over, and then decides” (1). But, contra the transcendence vision, the self doing this deliberation is not a separate entity from the processes that constitute it, the self’s choosing on the basis of reasons is consistent with its being caused to choose by (some of) its psychological states, and the possibility of predicting action based on the states and processes in the mind is not precluded.

We think that some non-reductive physicalist form of naturalism provides the best metaphysical theory and epistemological approach to understanding and describing the mind and agency. And we think such a theory can support a viable compatibilist theory of free will. For the purposes of this paper, however, we will not try to defend the truth of any specific theories of non-reductive physicalism or of compatibilism. Our goal here is to suggest that most ordinary people are *not* committed to a transcendence vision that conflicts with naturalism or compatibilism.

In order for ordinary people to have an understanding of mind, action, or free will that *conflicts* with a particular metaphysical or scientific theory, their understanding has to have substantive *content* that conflicts with that theory. We agree with Knobe that the ordinary understanding of mind and action does not have the same structure as a scientific theory, except in the weak sense that people often explain and predict actions, typically in terms of mental states, goals, and character traits. But this is not because most people have a *theory* that involves a transcendent self, non-causal reasons, or agent-causal powers that competes with the content of scientific theories of mind. Rather, we suspect that most people are “theory-lite.”

By “theory-lite” we simply mean that most people—at least before delving into philosophy or theology—do not have commitments to the underlying structure of the mind or the underlying causal processes that connect mental states to each other and to behavior. When people consider a sentence such as “John went to New York because he wanted to visit his sister,” they presumably think that John has desires and that these desires play some causal role in his going-to-New-York behavior. They will likely explain and predict John’s behavior in terms of his desires, and with increasing knowledge of John and his environment, they may offer increasingly precise explanations and predictions, with reference to increasingly precise mental states (e.g., reasons that explain why he took the bus rather than the train). And barring mitigating factors, they will think John acted freely and is responsible. But people do not thereby commit themselves to metaphysical beliefs about the nature of John’s mind, or the source of his desires or decision-making capacities, being either non-physical or physical, non-causal or (in)deterministically caused, supervenient on lower-level mechanisms or not.

Another way to put this point is that people have a relatively non-negotiable understanding of humans’ basic capacities to make choices and control their actions, but they have relatively negotiable or revisable beliefs about what underlies these capacities—that is, the metaphysical or scientific nature of the substance(s), processes, or sources of them.[[3]](#footnote-3)

Below we will suggest that the best explanation for people’s responses to our scenario and for results from other work in experimental philosophy supports this theory-lite interpretation of ordinary views, rather than a theory-laden view like the transcendence vision. But first we need to respond to an objection that might appear to stop us dead in our tracks.

Almost every culture and religion includes beliefs and practices about the self or soul being able to survive after bodily death and to exist outside the body and about agents being different in kind from other entities and animals. Recent research suggests support for “folk dualism” across cultures and beginning in young children (e.g., Bloom 2004; Bering 2006). Indeed, in work with Thomas Nadelhoffer, Nahmias finds that most people express a belief that humans have souls. For instance, in one sample 82% agreed that “Each person has a non-physical essence that makes that person unique,” and 61% agreed that “Human action can only be understood in terms of our souls and minds and not just in terms of our brains.”[[4]](#footnote-4) If the folk theory is substance dualism, then the naturalistic vision is not folksy.

However, belief in and talk of souls, the afterlife, out-of-body experiences, and human uniqueness is not evidence of a substance dualist theory of mind or a transcendence vision of free will. Such beliefs are themselves theory-lite. Different cultures and religions describe disembodied existence in different ways, and in most cases the self (soul or spirit) maintains a spatio-temporal form of some sort, often with quasi-physical properties. Few if any describe something like a Cartesian soul with no perceptual capacities, no clear way to be recognized or distinguished from others, and no spatial location (see Hodge 2008). The afterlife in traditional Judaism and Christianity involves resurrection of the body, not disembodied existence (see Murphy 2006). And even if some theological traditions are explicitly dualist, including those traditions’ libertarian descriptions of free will, it is quite common that people do not internalize the metaphysical tenets of their religion’s theology.

If we wanted to pin a philosophical theory of mind onto the folk, the best bet might be some form of functionalism, since functionalism allows that the mind or self can be instantiated in any underlying stuff that can maintain the interrelationships between experiences, memories, emotions, and thoughts that make a person the person she is. People seem more committed to the existence and relationships between certain mental categories than to any specific metaphysics of the mental (see also Phelan and Buckwalter, forthcoming).[[5]](#footnote-5)

In short, people’s understanding and talk of the ‘soul’ might best be described as a placeholder for whatever underlies the set of capacities humans have for thinking, feeling, and acting, and whatever distinctive features of an individual’s capacities make her the unique person she is. What fills this placeholder is, like many depictions of spirits and souls, vague and indistinct.

Most importantly for our purposes here, even if some cultures and people have a more specific conception of a non-natural soul that does not jibe with the naturalistic vision, this does not commit them to idea that free will or moral responsibility depend on such a soul. For instance, in the largely religious sample described in note 4, only 30% agreed with the statement, “If it turned out that people lacked non-physical (or immaterial) souls, then they would lack free will,” (with another third offering the neutral response), and 74% agreed with the statement, “People could have free will even if scientists discovered all of the laws that govern all human behavior.” Furthermore, Monroe & Malle (2010) found that people do not mention non-physical minds or souls when they describe free will, and in current work (Monroe et al., in preparation), they use experimental manipulations to show that people’s concept of free will is not tied to their concept of the soul. And Mele (2012) found that most people responded that agents can have free will and moral responsibility even if physicalism is true.[[6]](#footnote-6)

Finally, if people were committed to a dualistic or transcendent mind as grounding free will and responsibility, then they should be threatened by the possibility that their decisions and actions depend on, and can be predicted by, their brain processes. But, as we will now see, few people respond this way.

**2. My Brain Did It… and So Did I**

With a theory-lite understanding of mind and agency, people should not find some descriptions of naturalism threatening to free will or responsibility. To test this idea, we developed scenarios aimed at describing a possibility that rules out a transcendent self, or at least one that plays a causal role in actions, but that does not rule out a role for mental states and psychological processes. We cannot be sure that people are interpreting the scenarios in this way—one of the ongoing goals of experimental philosophy is to develop better methods to test how people are interpreting philosophical issues and which features of scenarios are driving their responses. We will discuss below some of the results that give us more information about people’s interpretations of our scenario.

This is the scenario participants read:

Recent brain scanning studies have shown that specific patterns of brain activity can be used to predict simple decisions several seconds before people are consciously aware of those decisions. Imagine that in the future brain scanning technology becomes much more advanced. Neuroscientists can use brain scanners to detect all the activity in a person’s brain and use that information to predict with 100% accuracy every single decision a person will make before the person is consciously aware of their decision. The neuroscientists cannot, however, do anything to change brain activity and hence they cannot directly influence thoughts and actions.  
  
Suppose that in the future a woman named Jill agrees, as part of a neuroscience experiment, to wear this brain scanner for a month (it is a lightweight cap). The neuroscientists are able to use real-time information about her brain activity to predict everything that Jill will think or decide, even before she is aware of these thoughts or decisions. However, they cannot alter her brain activity to change what she thinks and does.  
  
On election day, Jill is considering how she will vote for President and for Governor. Before she is aware of making any decisions, the neuroscientists can see, based on her brain activity, that she is about to decide to vote for Smith for President and Green for Governor. Just as the neuroscientists predicted, Jill votes for Smith for President and Green for Governor. As with her decisions to vote for Smith for President and Green for Governor, the neuroscientists are able to predict every decision Jill ends up making with 100% accuracy while she is wearing the scanner.  
  
Occasionally, Jill tries to trick the neuroscientists by changing her mind at the last second or by stopping herself from doing something that she just decided to do, but the neuroscientists predict these events as well. Indeed, these experiments confirm that all human mental activity is entirely based on brain activity such that everything that any human thinks or does could be predicted ahead of time based on their earlier brain activity.

The survey was completed by 147 students in critical thinking classes at Georgia State University. After excluding from analysis participants who missed one or both of the comprehension checks preceding the experimental questions or an attention check at the end, there were 122 participants whose responses were analyzed.[[7]](#footnote-7)

Participants were asked their level of (dis)agreement (on a 7-point scale with a midpoint of 4) with a series of statements about free will, responsibility, choice, and bypassing of mental states in three sets.[[8]](#footnote-8) First, there were statements about Jill’s specific acts of voting. Second, there were statements about Jill’s general actions performed during the experiment, all of which began “During the experiment, when the neuroscientists predicted what Jill did….” Third, there were statements about whether, “If this technology existed, then people…” would not have free will, be morally responsible, make choices, etc.

Responses to all of the statements strongly supported our predictions. For instance, 91% agreed that “Jill voted for Governor of her own free will” (with 50.8% strongly agreeing, mean = 6.09), and 90.2% agreed that “Jill was responsible for how she voted for Governor” (mean = 6.05).[[9]](#footnote-9) Similarly, 88.5% agreed that “During the experiment, when the neuroscientists predicted what Jill did, she had free will” (mean = 6.04), 94.3% agreed that she was responsible for what she did (mean = 6.14), 87.7% agreed that she made choices about what to do (mean = 6.00), and 86.9% agreed that what she did was “up to her” (mean = 5.95).

Only 12.3% agreed that “If this technology existed, then people would not have free will” (mean = 2.39), only 18.9% agreed that “If this technology existed, then people would not be morally responsible for their actions” (mean = 2.53), only 9.0% agreed that “If this technology existed, then people would not really make choices” (mean = 2.53), and only 11.5% agreed that “If this technology existed, then people would not truly deserve blame for their bad actions” (mean = 2.23).

In short, this detailed scenario describing perfect prediction of decisions based on brain activity did not lead people to conclude that free will is an illusion nor did it show that people think that the mind or self is something that transcends the causal order and makes decisions that cannot be predicted based on “the complex ways in which these states and processes interact” (Knobe, 1). Only a small minority of people take predictability based on brain processes to be incompatible with free will.

We suspect that most people are interpreting this scenario in a way that allows agents, like Jill, to have (conscious) mental states, to deliberate, and to act on reasons. That is, most people are not interpreting the fact that decisions are predicted by—even caused by—brain states to mean that decisions are not *also* caused by the agent’s reasons. Perhaps people are implicitly accepting some form of non-reductive physicalism or perhaps they are remaining theory-lite about how mental states and brain states are related to each other. In any case, most participants did *not* interpret the scenario to entail *bypassing*, or the view that mental states are causally irrelevant to behavior (see below). And on our view, that is the main reason they are not interpreting the scenario to rule out free will.

We disagree with Knobe about whether people think free will requires a transcendent self outside the realm of physical causality; however, we agree with him that a non-negotiable feature of people’s understanding of free will and agency is that agents cause their actions and that agents’ reasons have an effect on what they do. In our study we included statements to test whether people interpret this scenario to rule out either of these conditions. Consistent with the patterns of responses just described, 85.2% agreed that “Jill was the cause of how she voted for Governor” (mean = 5.80). Only 21.4% agreed that “During the experiment, when the neuroscientists predicted what Jill did, her reasons had no effect on what she did” (mean = 2.78). And only 15.5% agreed that “If this technology existed, then people’s reasons would have no effect on what they did” (mean = 2.56). Indeed, there was a high correlation between people’s responses to this bypassing statement (as well as others about whether this technology would show that people’s beliefs or desires have no effect on their actions) and their responses to statements about free will and responsibility (r = 0.59, p < .001).[[10]](#footnote-10)

These patterns of results suggest that only a minority of our participants (typically just one in five) find the possibility described in the scenario inconsistent with their understanding of free will and responsibility, and most of this minority group seem to be interpreting the scenario to involve bypassing of the agents’ mental states, perhaps because this minority group is employing a transcendent conception of the self. A significant majority of participants, however, do *not* see the naturalistic vision described in this scenario as suggesting bypassing nor do they take its potential truth to suggest that we would lack free will or responsibility.

Our scenario was a ‘concrete’ case, not an abstract case. That is, it included a specific agent making specific decisions. However, these decisions did not involve any emotionally laden actions.[[11]](#footnote-11) Hence, the ‘error theory’ offered in Nichols and Knobe (2007)—that high affect cases, such as theirs with a man murdering his family, lead people to attribute responsibility in deterministic universes—would not explain our results. Knobe no longer believes high affect is driving the concrete/abstract differences in attribution of free will or responsibility. In the previous chapter Knobe suggests instead that concrete scenarios may lead participants to apply their “default (transcendence-based) framework” and blame agents even in a deterministic universe (note 2). In our scenario, the agent is described concretely, and so, if Knobe’s recent explanation is correct, then participants deem her morally responsible because they take the uncaused self to cause her actions in accordance with her reasons. This explanation seems to us to require that most participants were not interpreting the scenario to rule out a transcendent self. We do not think this is the right interpretation of our results.

On the contrary, several results suggest that people are not simply rejecting the stipulations of the scenario as unrealistic nor are they responding without considering the ramifications of those stipulations. First, a significant majority of participants (80%) responded ‘yes’ when asked if they agreed with this statement: “It is possible for this technology to exist in the future.”[[12]](#footnote-12)

Second, when asked to explain why they thought the technology was or was not possible, no one mentioned free will or non-physical souls or minds. For instance, no participants said that the mind cannot be understood in terms of the brain or that we have minds or souls that cannot be seen or understood with such technology or whose activity cannot be predicted by what happens in the brain. Rather, only three said something about mind-reading being impossible, while most of the others who said the technology is not possible mentioned either the complexity of the brain, the high cost of such technology, or the moral or legal prescriptions people or governments would put on creating such technology.[[13]](#footnote-13)

Third, we have evidence that people do respond to the stipulations of such scenarios. In another scenario, almost identical to this one, the technology also allows the neuroscientists to alter Jill’s brain activity to change her decisions without her awareness. In that case, the vast majority of our 166 participants responded that, when her brain activity was altered, Jill lacks free will (80.2%) and responsibility (74.3%), and consistent with our view, the majority also responded that, in these cases, Jill does not cause her actions (74.3%) and her reasons do not play a role in her actions (59.4%). Hence, people do not take a “free will no matter what” position, and we have reason to believe that participants are responding to the specific features of the scenarios. For instance, they seem to be responding to the fact that the neuroscientists are indeed “reading” Jill’s mental states by seeing patterns of her brain activity, and they are expressing their belief that such mind-reading *without* manipulation is not a threat to freedom or responsibility.

Finally, we also asked people whether they agreed with this statement: “If this technology existed, it would show that each decision a person makes is caused by particular brain states.” The majority (65.6%) agreed with this statement, even though the scenario does not explicitly say this or entail it—it does conclude with the theory-lite claim that “these experiments confirm that all human mental activity is entirely based on brain activity.” Responses to the causal statement did not correlate with other responses (except for two of the bypassing statements)—that is, participants who disagreed with this statement were no more (or less) likely to disagree with statements about Jill’s or people’s free will or responsibility. Hence, most people seem to interpret this scenario to mean that Jill’s *reasons* are both caused by her brain states and cause her decisions. To the extent that people’s theory-lite view is being “filled in” by the scenario, then it is likely that they assume that the neuroscientists are predicting Jill’s behavior based upon those brain states which constitute her reasons.

It is, of course, difficult to know exactly how participants are interpreting relatively complex scenarios, including this one and many of the other scenarios used in experimental philosophy research, including some of the ones Knobe discusses. Much more research on these issues is required. However, by asking a variety of questions, including open-ended responses, and examining the patterns of responses, we think the best interpretation of our results is that most people have a theory-lite view of the mind and free will, one that is consistent with naturalism and that suggests no commitment to Knobe’s transcendence vision (and one that belies Harris’s assumptions about the ordinary view of free will).

We turn now to other results from experimental philosophy to argue that they too support this interpretation.

**3. The Causal Competition Principle**

Our brain-scan scenario does not explicitly describe determinism, though it suggests that prior brain activity is sufficient for (and predictive of) decisions and actions. It is possible that some people would respond differently if the scenario were more explicitly deterministic.

However, prior research by Nahmias and collaborators (2006, 2007, 2012) suggests that most people do not find determinism to threaten free will or responsibility, and of those who do, most do so because they misinterpret determinism to mean bypassing—that psychological states and processes play no role in action. Knobe suggests that these results, as well as his own, are better explained by assuming that most people have a transcendence vision of free will and, when they properly understand determinism, they take it to conflict with that vision.

Knobe suggests this 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.

That principle is very hard to dispute.[[14]](#footnote-14) But Knobe further suggests that this principle best explains the current body of research in experimental philosophy on free and responsible action—a claim that is easier to dispute. His interpretation of that data suggests that people go through this chain of reasoning:

Determinism 🡪 No production of behavior by agent 🡪 No free will or responsibility

Again, we agree with Knobe’s principle that constitutes the second step in this chain. But we disagree with his interpretation of the first step.

First of all, most people do *not* interpret determinism to mean that agents do not exist or that agents do not produce behavior. In Murray and Nahmias (2012), the majority of participants did *not* make the “bypassing mistake”—that is, they did not misunderstand determinism to mean that agents’ beliefs, desires, and decisions play no role in what they end up doing (and most did not make that mistake in our brain-scan scenario). Only in Nichols and Knobe’s abstract case did a significant majority of participants make the bypassing mistake. Knobe interprets these (minority) responses to suggest that people think that (a) determinism rules out reasons explanations, because (b) reasons explanations are incompatible with causal explanations, but since (c) reasons explanations are explanations in terms of beliefs and desires, (d) determinism entails bypassing of beliefs and desires (7-8). And Knobe suggests that the other participants (who constitute the majority, especially in concrete cases) do not draw this conclusion from determinism because they interpret the agent’s behavior as the result of the agent’s transcendent self and this interpretation leads them to neglect the alleged threat of determinism.

We think a simpler explanation of the body of results first recognizes that most people do not make the bypassing mistake, because they see that even if an agent’s mental states are fully determined by prior causes, that does *not* entail that the agent’s mental states, including her reasons, are causally inessential to her actions. And in concrete cases with descriptions of an agent making decisions for reasons, people are even more likely to see that the agent’s mental states play a role in her actions rather than being bypassed. A similar process seems to occur in most participants reading our brain-scan scenario.

Second, people who *do* make the bypassing mistake are likely doing so because they perceive there to be “causal competition” between the factors presented in the scenario and the agent’s mental states. We suggest this “causal competition” principle (CCP):

People will be reluctant to hold an agent responsible for behavior when they interpret her behavior as being fully caused by factors that do not include any of her reasons (or by processes that do not include any of her reasoning).

CCP is consistent with Knobe’s idea that people will sometimes interpret reasons explanations to be inconsistent with causal explanations. But we think this occurs specifically when behavior is seen as produced by causal factors that are not responsive to—that do not vary according to—an agent’s reasons.

CCP is similar to Björnsson and Persson’s (2012) view that people will hold agents morally responsible when the agents’ motivational structure is (1) such that holding them responsible will effectively alter or reinforce their behavior, (2) their motivational structure is a salient explanation of their behavior, (3) the behavior is such that it can be altered or reinforced, and (4) their motivational structure is morally significant (cf. Fischer and Ravizza 1998).

According to CCP, some descriptions of determinism are likely to lead people to think that agents are not free and responsible. This will occur, for instance, when people interpret the distant past as a complete cause of agents’ behavior such that it competes with the agents’ own mental states. Rather than taking agents’ reasons to be produced by (and responsive to) prior events and also as causes of decisions, people may take determinism to mean that everything is caused by the distant past, such that decisions must occur regardless of the agents’ reasons. For instance, most participants read Nichols and Knobe’s Universe A to mean that an agent will make certain decisions *no matter what*, such that the agent’s decisions would not vary even if she had different mental states. This may be because their description of Universe A is presented in direct contrast with Universe B, in which *human decisions* occur and these decisions do *not* have to happen the way they do, suggesting that in Universe A decisions do *have* to happen the way they do *no matter what* (see Murray and Nahmias 2012). We think the juxtaposition of Universe A and Universe B leads many people to think that a person’s actions have to happen the way they do regardless of her reasons, perhaps even that humans do not really reason or make decisions in Universe A (after all, participants are told that Universe B is just like Universe A but “the one exception is human decision making”).

Given that these descriptions of Universe A vs. B focus solely on this difference in decision making, we also do not find it surprising that fewer people make the bypassing mistake regarding processes like computer programs or involuntary behavior. Indeed, Knobe’s examples of non-voluntary actions, such as eye blinks and making facial expressions, are also cases where the behavior is not responsive to reasons. So, because the difference between Universes A and B is focused solely on human decision-making, there is no reason for people to interpret Universe A as presenting causal competition between past events and non-voluntary behavior or non-human processes.[[15]](#footnote-15)

CCP can also explain why reductionistic explanations of human decision-making are often taken to conflict with free will and responsibility. If an agent’s decisions and actions are presented as being completely caused by events at the neural level, with no mention of the agents’ reasons or reasoning, then it is easy to conclude that their decisions are not responsive to reasons. For instance, Nahmias, Coates, and Kvaran (2007) provided participants with scenarios that described the actions of agents as completely caused by either psychological states, such as thoughts, desires, and plans, or neurobiological states, such as chemical reactions and neural processes. They found that participants were significantly more likely to attribute free will and responsibility to agents whose actions were described as caused by their psychological states than those agents whose actions were described as caused by their neurobiological states, even though both cases described these proximal causes as being deterministically caused by earlier events.

Given that our brain-scan scenario also suggests causation by brain activity, one might wonder why most people do not make the bypassing mistake in that case. Even though the neuroscientists can predict Jill’s voting behavior with 100% accuracy based on her brain activity, people presumably interpret her voting behavior as responsive to reasons. People might think that if Jill were to have voted for Black for Governor (instead of Green), she would have had different reasons prior to casting her vote. While this counterfactual change of heart would also be predicted by the neuroscientists, they do not have the ability to alter her voting behavior—especially while holding her reasons constant. Consider instead the scenario we presented in which the neuroscientists could and did alter Jill’s voting behavior before she was consciously aware of her decision. In this scenario, it is likely that people felt the change in her vote was due to the neuroscientists’ alteration of her brain states (and perhaps their reasons) rather than her own reasons for casting a vote for Black.

This interpretation suggests avenues for future research. For instance, what alterations to the brain-scan scenario would lead most people to see it as undermining free will and determinism? CCP suggests the answer is any alteration that suggests a causal explanation for agents’ actions that competes with an explanation in terms of the agents’ relevant mental states. In the scenario in which the neuroscientists were able to use the scanner to alter Jill’s brain states, just the possibility of such manipulation slightly lowered ratings of her free will and responsibility (from 91% and 90.2% to 74.3% and 83.2%, respectively), even for decisions made without manipulation. For decisions that *were* manipulated, almost everyone responds that the agent *lacks* free will and responsibility (80.2% and 74.3%). In those cases, most people also judged that the agent’s decisions were not caused by her own reasons (59.4%). Such results are consistent with Knobe’s transcendence vision, but again there is no need to posit causal competition with a non-physical soul when a less theory-laden vision explains both these and our earlier results.

Another follow-up experiment would involve tweaking the brain-scan scenario to indicate that the prediction of decisions and actions is based on brain states regardless of what the agent experiences herself as thinking or wanting. Such a description would pull apart mental states from their neural bases and prioritize the neural processes as causes of action, activating CCP. Conversely, we could develop the scenario to make more explicit that Jill’s reasons for voting the way she did were completely caused by her other psychological states like beliefs and desires. If people agreed that Jill’s reasons could be completely caused in this way, it would suggest that the folk do not have a non-causal theory of reasons, as suggested by Knobe. He should predict that people would say that Jill’s reasons in such a voting scenario could be different even if all her other psychological states (e.g., beliefs, desires) were the same. We predict that most people would respond that Jill’s reasons could not be different holding fixed all of her her mental states prior to her decision.

Most people seem comfortable with the idea that our reasons and reasoning can be predicted based on brain states, even instantiated in brain activity. Problems arise when we are instead presented with a picture of a person’s neural processes “leading a life of their own,” unrelated to what the person consciously thinks, wants, or decides. This is the picture that we believe leads some people, including some neuroscientists (such as Harris), to think that the advance of neuroscience is challenging human agency and responsibility. This is the picture that makes some sense of the otherwise silly slogan: “my brain made me do it.”

**4. The Way Forward**

We will conclude by asking: why would anyone think that this vision of neuroscience—and this seemingly silly slogan—make any sense? In part, it is because the naturalistic vision remains, at this stage, a fuzzy sketch. At this early time in the development of the mind sciences, we still lack a theory to explain how mental processes, especially conscious ones, are ultimately constituted by neural processes (or physical processes more generally) and how to bridge explanations in terms of mental processes with explanations in terms of lower-level mechanisms. Hence, it is easy to see neuroscientific explanations as competing with, and preempting, ordinary psychological or reasons explanations, especially in contexts in which the neuroscientific explanations are presented as complete explanations of behavior. However, once we develop a naturalistic theory to explain how consciousness, thinking, and reasoning work, we may not experience the “explanatory gap” we currently experience when trying to understand how matter makes minds.[[16]](#footnote-16)

Of course, the naturalistic vision may get filled out in a way that does suggest causal competition between different levels or types of explanations for human action. Indeed, the sciences of the mind have already discovered various ways in which our understanding of our own minds is mistaken and needs to be replaced. [[17]](#footnote-17) Further discoveries may demonstrate that our ordinary conception of free will and responsibility is radically mistaken. However, for that to occur, the discoveries will have to provide some content that conflicts with the content of our ordinary understanding of these issues. We’ve suggested that such conflicts may be less likely to occur than Knobe and some scientists suggest, because, on the one hand, the content of the ordinary view of mind and agency is less specific and substantive than they suggest, and on the other hand, the sciences have not established a theory that conflicts with reasons explanations or establishes that conscious reasoning is not causal. Because the ordinary view is theory-lite, it is more flexible than often presumed. Because the naturalistic vision remains relatively fuzzy, it is not clear how it will be filled out. We see no reason to believe that the ordinary view and the naturalistic vision will not, in the end, overlap sufficiently to conclude that they are still “talking about the same things,” including free will and moral responsibility. We see no reason to conclude, as Knobe does, that a scientific theory of the mind and of free will requires abandoning our ordinary notions and replacing them with very different ones.

**References**

Bering, J. (2006). The folk psychology of souls. *Behavioral and Brain Sciences.* 29(5): 453-462.

Björnsson, G. and Persson, K. (2012). A unified empirical account of responsibility judgments. *Philosophy and Phenomenological Research*. doi: 10.1111/j.1933-1592.2012.00603.x

Bloom, P. (2004). *Descartes’ Baby: How the Science of Child Development Explains What Makes Us Human.* New York: Basic Books.

Cashmore, A. (2010). The Lucretian swerve: The biological basis of human behavior and the criminal justice system. *PNAS,* *107*(10): 4499-4504.

Coyne, J. (2012). Why you don’t really have free will. *USA Today* (1/1/2012).

Doris, J. (2002). *Lack of character: personality and moral behavior*. New York: Cambridge University Press.

Fischer, J.M. & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility.* Cambridge: Cambridge University Press.

Frankfurt, H. (1988). *The importance of what we care about*. Cambridge: Cambridge University Press.

Greene, J. & Cohen J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London B*. 359: 1775-1778.

Harris, S. (2012). *Free will*. New York: Free Press.

Hodge, K.M. (2008). Descartes’ mistake: How afterlife beliefs challenge the assumption that humans are intuitive Cartesian substance dualists. *Cognition and Culture*. 8: 387-415.

Mele, A. (2012). Another scientific threat to free will? *The Monist.* 95: 422-440.

Monroe, A. E., Malle, B. D., & Dillon, K. D. (in preparation). Is the soul to blame?

Monroe, A. E. & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people’s folk concept of free will. *Review of Philosophy and Psychology*. 1(2): 211-224.

Montague, R. (2008). Free will. *Current Biology*. 18(14): R584-585.

Murphy, Nancey. (2006). *Bodies and Souls, or Spirited Bodies?* New York: Cambridge University Press.

Murray, D. & Nahmias, E. (2012). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*. doi: 10.1111/j.1933-1592.2012.00609.x

Nahmias, E. (forthcoming). Is free will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (ed.), *Moral Psychology, vol. 4: Freedom and Responsibility*. Cambridge: MIT Press.

Nahmias, E. (2007). Autonomous agency and social psychology. In M. Marraffa, M. Caro & F. Ferretti (Eds.), *Cartographies of the Mind: Philosophy and Psychology in Intersection*. Dordrecht: Springer.

Nahmias, E., Coates, J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: experiments on folk intuitions. *Midwest Studies in Philosophy*. 31(1): 214-232.

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

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs.* 41(4): 663-685.

Phelan, M. & Buckwalter, W. (in preparation). Analytic functionalism and mental state attribution.

Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson’s.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., and Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language*. 25(3): 346-358.

Stoljar, D. (2009). Physicalism. *Stanford Encyclopedia of Philosophy.*

Tancredi, L. (2007). The neuroscience of ‘free will.’ *Behavioral Sciences and the Law.* 25: 295-308.

Tononi G. (2008). Consciousness as integrated information: a provisional manifesto. *Biological Bulletin*, 215: 216-242.

Wolf, S. (1990). *Freedom within reason*. New York: Oxford University Press.

Young, L. (in preparation). The Conflicted Self does not Cause its Own Actions.

1. We are grateful to Shane Reuter and Jason Shepard for helpful discussions on these topics and collaboration on the new study described in this chapter and to Al Mele and Josh Knobe for helpful comments on earlier drafts. [↑](#footnote-ref-1)
2. Similar claims by scientists who argue that neuroscience threatens free will are widespread and receive much attention in the popular press. See, for instance, Greene and Cohen (2004), Tancredi (2007), Montague (2008), Cashmore (2010), and Coyne (2012). For responses to these arguments, see Nahmias (forthcoming). [↑](#footnote-ref-2)
3. Yet another way to put this is to say that people don’t have strong views about the underlying metaphysical or scientific nature of the sorts of things that are part of what Sellars called the “manifest image,” things such as colors, money, dogs, beliefs, agents, or solidity. People may be surprised to find out that solid objects are composed of mostly empty space, but that does not change most of their beliefs about, or talk of, solid objects, much less the way they interact with solid objects. It does not lead people to believe or say that solidity is an illusion. [↑](#footnote-ref-3)
4. These results come from 330 participants from Qualtrics who are representative of the demographics of the general U.S. population. Knobe also suggests that the results of Sarkissian et al. (2010) provide evidence for cross-cultural beliefs in incompatibilist free will. However, those experiments show that in every culture, between one-third and one-quarter of participants offered the minority response, and we think that it is a mistake to interpret the majority response to the abstract version of Universe A as indicative of a commitment to dualism or to incompatibilism (see below). [↑](#footnote-ref-4)
5. It is helpful to remember that even though Gilbert Ryle (1949) calls Descartes’ view the “Official Doctrine,” he then argues at length that our ordinary talk and beliefs about mental phenomena suggest a behaviorist (or perhaps better, functionalist) folk theory of mind, whereas substance dualism is driven by philosophical mistakes. While people may not think of mental phenomena as physical, that does not thereby mean that they think of mental phenomena as *non-*physical. [↑](#footnote-ref-5)
6. For instance, 73% agreed that agents could have free will after reading a scenario that included this description: “In 2019, scientists finally prove that everything in the universe is physical and that what we refer to as ‘minds’ are actually brains at work” (Mele 2012: 430). Participants were not just responding that people have free will no matter what, since 79% said that agents would not have free will in a scenario describing a compliance drug. [↑](#footnote-ref-6)
7. Though college students are not particularly representative of the general U.S. (much less global) population, GSU students are more representative of the religious, racial, and socio-economic status of the U.S. population than most university populations. Nonetheless, 25% of this sample was non-religious, which is roughly double the general population (we have not yet tested whether religious affiliation or religiosity correlates with any of the responses). Comprehension checks asked about whom Jill voted for and the attention check asks participants to answer “None of the above” in response to a long statement. [↑](#footnote-ref-7)
8. Statements were randomized within each of these three sets. [↑](#footnote-ref-8)
9. Responses were almost identical to statements about her voting for President. The two decisions were included because another study used a scenario in which the neuroscientists could alter her decisions before she was aware of making them, and we compared responses to statements about a decision that was not altered with one that was altered (see below). [↑](#footnote-ref-9)
10. For this correlation, we used a composite bypassing score (Cronbach’s Alpha = .870) and a composite free will and moral responsibility score (Cronbach’s Alpha = .703). [↑](#footnote-ref-10)
11. Presumably, mention of elections with no party names or recognizable political names is not stirring up participants’ emotions. [↑](#footnote-ref-11)
12. Furthermore, responses to this question about the possibility of this technology did *not* correlate with responses to any statements about the consequences of this technology for free will or responsibility. It correlated with only one other statement: “If this technology existed, then people’s reasons would have no effect on what they did” (r = .196, p = .031). [↑](#footnote-ref-12)
13. For what it’s worth, we would respond that such technology is, in practice, impossible, since we think no technology will be able to both measure the incredibly complex activity of some neurobiological processes and simultaneously calculate the outcomes of such activity so as to predict *all* behavior with 100% accuracy. Assuming that conscious mental processes supervene on neural processes, then it is likely that some conscious deliberation processes cannot be predicted perfectly *prior* to their, and hence the relevant neural processes, themselves occurring (see Nahmias [forthcoming] for discussion of how this point provides one response to the common claim that Libet-style results challenge free will). Of course, these points, like the ones raised by our participants, are consistent with the truth of naturalism, of determinism, and of the compatibility of free will with both. [↑](#footnote-ref-13)
14. Note, however, that it is not very informative, since (a) on its own, it tells us nothing about *when* people think an action is or is not produced by an agent, and (b) it tells us nothing about the many circumstances in which people mitigate an agent’s responsibility even though they also think she produced the behavior—e.g., ignorance, coercion compulsion, addiction. If one argues that these cases are best understood as cases where the agent does not really produce the behavior, then one seems to be adopting a compatibilist theory, such as the deep self view, which identifies the *responsible* agent with some subset of the agent’s overall psychology. [↑](#footnote-ref-14)
15. This interpretation would be strengthened if the studies Knobe describes included the contrastive description of decision-making in Universe B, but we cannot tell whether or not he included the description of Universe B in these studies. CCP may also help to explain recent evidence that suggests that participants are less willing to hold an agent responsible for an action about which she is conflicted (Young, in preparation). For example, an agent who wrinkles her face at her gay roommate having sex, even though she supports gay rights, is judged to be less responsible for wrinkling her face than an agent described as opposing or being indifferent towards gay rights. That is to say, participants judge those agents with psychological states that conflict with ‘endorsed’ psychological states (reasons) to have less control of their resulting behavior compared to agents whose behavior does not conflict with endorsed psychological states. CCP explains the participants’ reluctance to hold such agents responsible for their behavior: their behavior is produced states that conflict with their reasons. [↑](#footnote-ref-15)
16. Just as Copernicus provided good reasons to accept the heliocentric model, we have good reason to accept the naturalistic vision, but just as the Copernican model did not make sense of our experience of the earth being unmoving until Galileo’s theory of inertia explained it, we lack a theory to help us understand how brain processes can explain our conscious experiences (for a valiant attempt, see Tononi 2008). [↑](#footnote-ref-16)
17. For instance, see Doris (2002) and Nahmias (2007) for discussions of the ways situationist social psychology raises challenges to our ordinary understanding of character traits and responsibility. [↑](#footnote-ref-17)