Your Brain as the Source of Free Will Worth Wanting:
Understanding Free Will in the Age of Neuroscience

by Eddy Nahmias, Georgia State University

for *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*
edited by Gregg Caruso and Owen Flanagan (Oxford University Press)
[Prepublication draft.]

1. Three Reactions to Neuro-naturalism

Imagine you have an important decision to make about which of two job offers to accept. You must decide by 5:00 pm. The offers, A and B, have various competing attractions and drawbacks. Currently, there is no answer to the question of what you will decide. But there needs to be by 5:00 (option C of picking neither and being unemployed is not on the table). You have several more hours to consider your reasons for each option, to discuss them with friends and family, to imagine how your life will go if you choose A and how it will go if you choose B. This feels like an existential choice, since your future depends upon it: your life will be significantly different depending on what you decide.[1] Some of these differences are evident to you—they are the ones you imagine and weigh against each other—others are unknowable, but you cannot do anything about those. You also realize that some of your reasons, and how important they seem to you, are influenced by factors you don't know about, some of which you wouldn't want to influence you. But you do the best you can to consider what you think *is* most relevant and to decide based on what really matters to you. Of course, this means you are also making some decisions along the way about what matters to you and how much. As the day wears on, you find yourself leaning towards option B. It's only 4:30, so you continue to deliberate, testing your reasons and your feelings about B, letting yourself plump for A to see how it feels. But just before 5:00, you make your decision final by sending an email to A to decline and by calling B to accept.

Not all decisions are like this, of course. Most are not so existential (e.g., choosing between soup and salad for lunch), many are made without such extensive conscious deliberation or with less rational consideration, and alas, many are made without feeling as confident about what to do by the time the decision must be made. I hope you will fill in the example sketched above with an *actual* decision you've made that has these features: an important choice for which you imagined various alternative outcomes, evaluated your reasons and feelings, and eventually came to a relatively confident decision about what to do. (I'll wait here while you think about it.)

[1] I recently advised a student who had to decide whether or not to major in philosophy before he registered for classes. While this decision is not as significant as the one Sartre describes of the young man deciding between joining the Resistance and staying to care for his mother, the student I advised described his decision as "existential."

These decisions seem to represent one paradigm of free and responsible agency. But what if I told you that all of the mental processes involved in making your choice—imagining the options, evaluating them, making the decision—*all* of these processes happened … *in your brain*? Indeed, each of those mental processes just *is* (or is *realized in*) a complex set of neural processes which causally interact in accord with the laws of nature. Call this thesis about the relation of mental processes and neural processes "neuro-naturalism."[2]

If you are like me (and forgive the pun), your mind is not blown by this assertion of neuro-naturalism. It may instead seem banal, though at the same time a bit mysterious, since we do not yet understand how neural processes could achieve all of these remarkable conscious and rational decision-making tasks. I think my reaction is a common one, at least among contemporary educated people, and below I'll provide some evidence for this. I will call it the "natural reaction" to neuro-naturalism.

Some people, however, find neuro-naturalism patently absurd or impossible; for instance, they are committed to a dualistic conception of the mind and free will. They do not accept that *mere* neurobiological activity could explain consciousness, imagination, or decision-making, and hence they resist the possibility of neuro-naturalism, and take its assertion to be a threat to free will. Call this the "dualist reaction."

Finally, others (call them "pessimists") take the neuro-naturalist understanding of our mind and agency to be angst-inducing. On the one hand, they take most people to be wedded to the dualist *conception* of mind and agency, but on the other hand, they accept the truth of neuro-naturalism, typically a reductionistic brand of it. So, the pessimists think that if people could be induced to get their head out of the sand, the truth would blind them. Most people, having a dualist understanding of self and free will, would fight to put their heads back in the sand or only painfully come to accept the truth. To be fair, most pessimists think that, even if the truth initially causes some angst, it will also rid us of some harmful illusions and, in the end, be beneficial.[3]

My goal in this chapter is to provide some diagnoses of these different reactions to neuro-naturalism, and provide some reasons to think that the natural reaction is both common and correct. Focusing on free will, I will offer reasons to think that a neuro-naturalistic understanding of human nature does not take away the ground (or grounding) that supports most of our cherished beliefs about ourselves, any more than Copernicus' shifting the earth from the center of the universe took away the ground that supports us. It did take Galileo's theory of inertia to make sense of how the earth can be flying through space while we feel unmoving, supplemented

[2] Neuro-naturalism, as I'll use it, is meant to be compatible with various forms of physicalism in philosophy of mind, including both non-reductive and reductive varieties (Stoljar 2009). However, neuro-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, nor to a reductionistic epistemological thesis that says the best explanations are always those offered by lower-level sciences (e.g., physics or neuroscience).

[3] The 'pessimist' label is drawn from P.F. Strawson (1962), whose views I hope to reflect, if only dimly, here (as I do with Daniel Dennett's views in his 1984 book, whose subtitle I use in my title). Strawson's use of 'pessimist' to refer to incompatibilists about free will and determinism is inapt in some of the same ways my use is, since many free will skeptics are *optimistic* about the benefits of our giving up outmoded views of free will (e.g., Pereboom (2014), and chapters in this volume by Pereboom and Caruso (Ch.11) and Focquaert, Glenn, and Raine (Ch.13)).

by his helpful analogy with our feeling unmoving in the hull of a smooth-sailing, fast-moving ship. But with that explanation of our experience in place, most people could grow up learning the Copernican theory without existential angst. We *experience* the earth as unmoving, and we need Galileo's *theory* to make sense of that experience. But once that experience is accounted for, it was those most committed (for scientific, philosophical, or religious reasons) to the competing Ptolemaic or Aristotelian *theory* who felt the most angst about the Copernican revolution.[4]

Similarly, dualists and reductionists, committed to their competing theories, tend to think neuro-naturalism conflicts with people's self-conception. But, I will suggest, most people are 'theory-lite' and amenable to whatever metaphysics makes sense of what matters to them. We do not yet have a theory like Galileo's to explain how neural activity can explain our conscious experiences. However, I predict that such a theory will have to make sense of how those neural processes involved in our deliberating—for instance, about what job offer to take—are crucial causes of our decisions about what to do. I will suggest below that interventionist theories of causation offer the best way to see this. The neuro-naturalistic picture has already begun to seep into the public consciousness, and many people have the natural reaction, because they seem to assume that a future theory will be able to make sense of our experiences *within* the neuro-naturalistic picture. Hence, their lack of angst. If and when such a theory actually emerges, then even though it will establish that there is "nothing" more to us than our complex brains and bodies existing in a physical world, governed by the laws of nature, it likely will also make sense of how we can have a type of free will that can ground our being unique, creative, unpredictable, imaginative, autonomous agents who are the sources of our actions.

2. A Transparent Bottleneck … or Nexus

In a much discussed piece, Joshua Greene and Jonathan Cohen (2004) argue that neuroscience has vindicated a reductionistic form of neuroscience that will provide a window for people to see the threats it poses. They assume that people have deeply-held implicit or explicit beliefs about the mind as a non-physical entity and about free will as a libertarian power to make decisions ungoverned by natural laws. But Greene and Cohen think that the opaque metaphysical theses of naturalism and determinism are not vivid enough to pull people's heads out of the sand

[4] God has the power to move heaven and earth, and religions eventually moved their conception of earth to its actual place in the cosmos. Similarly, an all-powerful God would have the power to create persons in fully physical form. Religions can, and have, imagined their God or gods creating humans without non-physical souls but with all the good stuff typically ascribed to souls, such as consciousness, identity, free will, a moral sense, even eternal life. The point is that religious belief is not wedded to dualistic belief, and dualistic religious tenets are not strong evidence of deep dualistic *intuitions* (or a dualistic folk psychology). Furthermore, the concept of a non-physical soul does not *explain* how we have consciousness, free will, etc. Rather, it serves as a placeholder essence that somehow has these properties without explaining them. People's use of terms like 'soul' and 'mind' often occurs in a causal-explanatory framework and only rarely refer to any alleged non-physical attributes of either.

so they can see how their dualist beliefs are challenged by these theses, and this explains why these theses have yet to shake up our moral and legal systems.

Neuroscience, on the other hand, will illuminate the mechanisms of decision-making in a way that will challenge people's beliefs and hence challenge our moral and legal practices (notably, our retributive punitive system): "neuroscience holds the promise of turning the black box of the mind into a *transparent bottleneck* … your brain serves as a bottleneck for all the forces spread throughout the universe of your past that affect who you are and what you do. Moreover, this bottleneck contains the events that are, intuitively, most critical for moral and legal responsibility" (p. 1781).

That's one way to look at it. But we can also flip this image on its head to recognize that neuroscience will open up the black box of the mind to illuminate how the very processes that we take to be critical to decision-making actually work. We can recognize the "transparent bottleneck" of the brain as the complex *nexus* that brings together a remarkable amount of information from both the past (including genes, upbringing, and learning) and the present (including external stimuli and internal beliefs, desires, goals, etc.). This nexus then serves as the source of the causal activity that integrates (some of) this information as we make decisions. These integrative processes, according to the natural reaction to the neuro-naturalist picture, will indeed be the ones most critical for responsibility, such as our consciously weighing options and reasoning about what to do. Because each nexus of neural activity—that is, each of our brains—is the site of a unique causal history, this picture also helps to explain why each of us, along with our subjective experiences, is unique.

On the neuro-naturalist view, neural activity also has to explain the existence of our conscious experiences as we make these decisions. Again, while we lack a theory that explains all the features of conscious experiences, especially its subjective or qualitative features, assuming (as I am here) that such a theory is forthcoming, it will presumably illuminate where and how the brain represents our conscious imagining of future options (like job offers A and B), of likely outcomes of choosing those options, and of evaluations of those outcomes, including emotional reactions to them. While the role of consciousness in agency is contentious (e.g., Levy 2014, Caruso 2012), it is plausible that conscious processes allow integration of a wide range of information, which seems crucial for the sort of imagination and evaluation described here (see Nahmias forthcoming and Sripada forthcoming for discussion of the role of imagination and prospection in free will and of brain regions likely responsible for such processes).

The modern mind sciences have, of course, discovered that our decision-making is subject to external stimuli and non-conscious internal states that can lead us to make sub-optimal choices, which we may then rationalize after the fact (see, e.g., Nahmias 2007). We are also learning that some genes and/or early experiences can have significant influences on how our brains are 'wired' and hence on our decisions, and in some cases, the result is 'faulty wiring' and bad decisions. If neuro-naturalism entailed that we are somehow unable to recognize genuine reasons for action or unable to control action in light of such reasons, then pessimism would be warranted. But for now, let us assume that these specific empirical challenges to our capacities

for rational decision-making and self-control do not universalize—that is, in many cases we are 'wired' in a way that *explains* the proper functioning of these capacities rather than explaining them *away*. In that case, we can examine the less radical neuro-naturalistic thesis that simply says that our conscious deliberation and rational decision-making, to whatever extent we actually possess them, are carried out by neurobiological processes (see Nahmias 2014 and Mele 2009 for responses to empirical evidence presented as challenging any causal role for conscious or rational processes).

To help us understand the neuro-naturalistic possibility, Greene and Cohen ask us to imagine a time in the future when "we may have extremely high resolution scanners that can simultaneously track the neural activity and connectivity of every neuron in a human brain, along with computers and software that can analyse and organize these data" (p. 1781). They ask you "to imagine watching a film of your brain choosing between soup and salad" for lunch. I will ask us instead to consider what we would see if we watched what occurs in our brains as we make the sort of complex decision I described above (italics indicate where I have altered their text accordingly):

> The analysis software highlights the neurons pushing for *offer A* in red and the neurons pushing for *offer B* in blue *(i.e., the neuronal processes that realize your imagining the pros and cons of each job offer).* You zoom in and slow down the film, allowing yourself to trace the cause-and-effect relationships between individual neurons—the mind's clockwork revealed in arbitrary detail. *After examining the neuronal processes involved in the extended and complicated mappings of the conscious deliberations as you imagined various consequences of, and reasons for, each choice,* you find the tipping-point moment, at which the blue neurons in your prefrontal cortex "out-fire" the red neurons, seizing control of your pre-motor cortex and causing you to *send the email to reject offer A*. (2004: p. 1781; compare Harris 2012: pp. 10-11)

Notice that it is awkward and radically incomplete to try to describe in these terms the astronomical complexity of the neural activity that would actually have to be captured and analyzed to illuminate what occurs as we spend extended periods of time considering an important decision. If we try to consider the complexity of what we would see occurring in the "bottleneck" of our brains as we make such decisions, we would see a process unfolding in the nexus of our brain over time and space, and we would not "shrink under this scrutiny to an extensionless point" (to repurpose Nagel's memorable phrase in 1979, p. 35). Trying to imagine a more complete account of such complex decisions is important if we aim to diagnose the different reactions to the neuro-naturalism that this futuristic brain-scanning is supposed to illuminate. We need to avoid selling our brains short. If neuro-naturalism is true, then the "film" of your deliberations and decision about which job to take will not really be reducible to mere images of blue and red neurons.

3. Avoiding the Bypassing Intuition

Are pessimists like Greene and Cohen correct to predict that most people will understand neuro-naturalism as conflicting with free will and challenging our moral and legal practices? I will suggest they are not. But first, let's consider why these pessimists assume there is a conflict. I suspect some neuroscientists are especially prone to see a conflict between their reductionistic methodology, with its focus on the mechanisms causing human and animal behavior, and folk psychological explanations of behavior that do not refer to such mechanisms. They might also think ordinary people, like scientists, have a substantive theory about the way mind and agency work, a dualist one. And these scientists are especially likely to recognize the lack of a scientific explanation of consciousness, like the pre-Galileo theorist who sees the lack of explanation for our experiences within the Copernican theory. As such, when these neuroscientists assume that neuroscientific explanations can explain and predict all human behavior, they may conclude that the unexplained conscious features of our mental life have no causal role to play—they are *bypassed*.[5]

But most people—at least those who do not delve into science, philosophy, or theology—do not have such substantive theories about how the mind or agency work. And the less substantive or specific their commitments to the underlying structure of the mind or the underlying causal processes that connect mental states to each other and to behavior, the less substance there is to be falsified by metaphysical or scientific theories. If people are 'theory-lite' in this way, then while they may have a relatively non-negotiable understanding of humans' basic capacities to make choices and control their actions, they may have relatively negotiable or revisable beliefs about what actually underlies or explains these capacities—that is, the metaphysical or scientific nature of the substance(s), processes, or sources of them. If so, then we should predict that people are not committed to a dualist understanding of free will that conflicts with neuro-naturalism, and hence they may not have the reaction predicted by Greene and Cohen to the possibility of seeing all the neural processes responsible for decision-making, at least so long as those processes are described as the neural instantiation of the mental processes "that are,

[5] For examples of other pessimist neuroscientists and psychologists, see Nahmias 2014. However, not all neuroscientists take the pessimistic view towards discoveries of what the brain does during decision-making. Consider a recent study that seems to bring to life the fictional one Greene and Cohen describe by finding the neural activity associated with the 'tipping point' in the brain when people made decisions about where to focus their attention (Gmeindl et al. 2016). The researchers found the activity (in specific areas of prefrontal cortex) that built up starting about three seconds before people shifted their attention, likely with their awareness of a decision to shift occurring during that buildup of activity. Like other fMRI studies using multivoxel pattern analysis (MVPA), the mapping had to be individualized to each participant's unique brain. The researchers do not take their approach as challenging free will, but instead as helping to discover how it works. In a media report titled "What Free Will Looks Like in the Brain" (seemingly an oxymoron for dualists and pessimists), one researcher says, the aim of the study is to "peek into people's brain and find out how we make choices … and what parts of the brain are involved in free will." Another says, "that by devising a way to detect brain events that are otherwise invisible—that is, a kind of high-tech 'mind reading'—we uncovered important information about what may be the neural underpinnings of free will." (http://www.eurekalert.org/pub_releases/2016-07/jhu-wfw071316.php)

intuitively, most critical for moral and legal responsibility." That is, as long as they are not led to believe that a neuro-naturalist picture entails bypassing of these critical processes.

In Nahmias, Shepard, and Reuter (2014), we presented people with a neuro-prediction scenario inspired by the ones that pessimists like Greene and Cohen and Sam Harris predict will lead people to see the threat to free will posed by neuro-naturalism.[6] Our scenario first describes future technology: "Neuroscientists can use brain scanners to detect all the activity in a person's brain and use that information to detect the activity that causes decisions and predict with 100% accuracy every single decision a person will make before the person is consciously aware of their decision." It states that in the future a woman named Jill agrees to wear the scanner for a month, during which time the neuroscientists are able to predict all of her decisions, even when she is trying to trick them, and including decisions about whom to vote for in an election. And it ends with a statement meant to suggest neuro-naturalism, in one version stating, "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," and in another, using the phrase, "all human mental activity just *is* brain activity" (see Nahmias et al. 2014 for details of studies and results).

When asked whether it is possible for such technology to exist in the future, 80% said it was. Pessimists, it would seem, should predict that many more people would reject the possibility of this technology, since a dualist or libertarian should reject the possibility of obtaining from physical information complete information about a person's (non-physical) mental processes during decision-making or the possibility of decisions being fully caused by prior neural activity, or both. In fact, almost none of our participants explained their responses to this possibility question with any mention of free will, souls, or the impossibility of understanding or predicting decisions based on brain processes.[7]

Furthermore, across a range of questions, 75-90% of participants responded that the technology would not conflict with free will or moral responsibility and that Jill was free and responsible while having all her decisions perfectly predicted by the neuroscientists. They do not see the threat predicted by pessimists. The minority who said the technology would threaten free

[6] Harris writes: "Imagine a perfect neuroimaging device that would allow us to detect and interpret the subtlest changes in brain function…. the experimenters knew what you would think and do just before you did it. You would, of course, continue to feel free in every present moment, but the fact that someone else could report what you were about to think and do would expose this feeling for what it is: an illusion" (2012, pp. 10-11).

[7] Instead, most who said it was possible referenced the remarkable advances of science and technology or the fact that our mental activity all occurs in our brains, while for the 20% who said it was impossible referenced the likely technological glitches or political and ethical resistance to developing such technology. Granted, our participants were college students with at least some scientific background (and while quite diverse at my institution, still less religious than the general population). However, this actually supports my overall view, since I assume that dualist *beliefs* (and avowals) are prevalent, at least in Judeo-Christian cultures, yet also revisable without too much resistance, as long as our folk psychological explanations are not being undercut. 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 or talk of mental phenomena in physical terms, that does not mean that they think of mental phenomena as *non*-physical.

will also expressed bypassing intuitions—e.g., agreeing that it would mean that people's reasons have no effect on what they do. The majority, however, did not express such bypassing intuitions. Instead, these participants seemed to be assuming that what the brain scanners are detecting (e.g., in Jill) is precisely the neural activity that instantiates the reasoning processes (e.g., as Jill considers what to do). That is, most people seem to interpret this scenario to mean that Jill's reasons and reasoning 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 that constitute her freely deciding what to do. They might also be implicitly assuming a post-Galileo theory of mind has been discovered, a future neuroscience that has figured out how "mere" neural activity can explain, rather than explain away, our imagining and evaluating future options and how those processes have the right sort of causal influence on our decisions.[8]

Supposing most people are theory-lite in this way such that a common response to neuro-naturalism is what I am calling the natural reaction, we might then wonder whether this is the *correct* reaction—whether there is a way to make sense of decision-making, even free will, in a neuro-naturalistic framework. I will now sketch a positive answer to that question.

4. Causal Sourcehood in a Neuro-Naturalistic Framework

Suppose that ordinary people, like scientists, think about causation in roughly the way suggested by interventionist theories of causation (see Sloman 2005, Lagnado et al. 2013). On this view, to know whether one event X causes another Y, we consider what would happen to Y if X (and nothing else) were different in various ways. More precisely, we consider interventions on the value of X, while controlling for the other causal influences on Y, and we see what happens to the value of Y (for details, see Woodward 2003). Furthermore, we can compare the relative strength of causal influences on an outcome by seeing which has a more causally invariant relation with that outcome. For instance, consider two genes, W and X, that influence phenotypic trait Y. W has a stronger causal invariance relation with Y than X just in case:

> (1) holding fixed relevant background conditions C, interventions on the value of W cause specific variations on the value of Y more so than interventions on X—e.g., holding fixed the rest of the organism's genome and environment, mutations of W influence the expression of trait Y more so than mutations of X do.
> (2) the causal relationship between W and Y remains across a wider range of relevant changes to background conditions than does the relationship between X and Y—e.g., the influence of gene W on trait Y remains across a wider range of changes in other genes in

[8] Another interpretation of these results is that people are so committed to a dualist or libertarian view of decision-making that they simply reject the stipulations of the scenario once they start thinking about a human who is making decisions (see Rose et al. 2015). Further research is required to test this alternative theory. For further experimental work suggesting that most people do not have dualist intuitions about free will, see Monroe and Malle 2010 and Mele 2012.

the organism or changes in the organism's environment than does the influence of gene X on Y. (see Deery and Nahmias forthcoming)

If we are looking for the *causal source* of a particular outcome, we would look for the causal influences of that outcome that have the strongest causal invariance relation with the outcome.[9] Goal-directed causes will typically have a strong causal invariance relation with their effects, since they will lead to adjustments in response to varying conditions in order to bring about the desired outcome. For instance, whether Romeo ends up kissing Juliet (variable K) will have a strong causal invariance relation with his desiring to kiss her (variable J), since (1) interventions on J (e.g., such that he desires to kiss Rosaline instead) would lead to very different outcomes than K, and (2) J will lead to K across many alterations of background conditions—Romeo's desire will lead him to reach Juliet despite a range of obstacles, such as walls to scale (see Lombrozo 2010, who uses the Romeo example from William James). Similarly, my goal G of having a challenging job might have a stronger causal invariance relation with my decision to accept offer B than other factors. (1) Holding fixed the actual circumstances C, varying G's value (e.g., by considering cases where I care less about the challenges) would influence my decision more than varying the value of any other factor. And (2) holding fixed G, it influences my decision more than other factors across the widest range of relevant changes to the circumstances C (e.g., no other causal factors lead to the same decision while altering conditions such as the relative salaries of the two jobs).

Now, imagine we're looking at Greene and Cohen's film of the neural activity as I'm making my decision, and allow me to oversimplify (though less so than they do). We're assuming a version of neuro-naturalism that does not eliminate psychological variables, so there must be various complex neural processes that realize the factors influencing my final decision, from ones I knew about and recognize ("Oh, there's goal G") to ones I did not know about ("Ah, now I can see the influence of the guilt I felt when I thought about moving away from my mother"), along with many other neural variables that influence my decision, some realizing psychological variables, many not. We could not do the actual interventions on this particular decision to test the relative strength of causal relationships, but the interventionist theory does not require that such interventions are, or can be, done (see Woodward 2003). Presumably, the neuroscientists had to do many experimental interventions to discover which neural pathways are relevant to which behavioral outcomes (including verbal reports about experiences while carrying out tasks). With futuristic optogenetic technology, perhaps they could intervene on specific neural processes to test various effects. Here, let us assume that for some decisions like

[9] We can also consider causal sourcehood as coming in degrees, such that we can say that W is the source of trait Y more than X is. On this view, much depends on how we are understanding the relevant outcome. For instance, my parents' conceiving me at the particular time they did (event X) may be the causal source of my existing (rather than not existing or rather than some other person existing). But, holding fixed my existence, event X is not the causal source of my deciding on job offer B (rather than A), at least not if we are considering X in relation to a variable such as my considering B to be a more challenging job than A, which has a stronger invariance relation with that particular decision (see below).

the one about which job to take, some of the variables with the strongest invariance relations to my decision are the very ones that I considered important as I deliberated (like goal G), the ones I am now observing in "neural form." If so, two important consequences follow regarding the causal source of my decisions.

First, in many cases psychological variables, such as my conscious imagining of options, will have a stronger invariance relation to my decision than neural variables, *even the ones that realize those psychological variables*, and hence they are plausibly understood to be the causal sources of my decision. As argued by Campbell (2009), List and Menzies (2010), and Woodward (2015) against the causal exclusion argument, interventionism suggests that psychological variables (e.g., beliefs or intentions) can be picked out as the cause of effects (such as decisions or actions) over the neural variables that realize them (or on which the psychological variables supervene). This is because (at least plausibly) the psychological variables could be realized by different neural variables, so interventions on the neural variables might not alter the effects, whereas interventions on the psychological variables would. For instance, holding fixed relevant background conditions, if my goal G is realized by neural variable N1, an intervention on N1 (replacing it with, say, N2, another state that can realize G) would *not* alter my decision for job B. Conversely, an intervention on G would. This argument does not require a commitment to mental states being multiply realized by computers, alien minds, or anything else besides brains (though it is consistent with that possibility). Rather, it only requires that at least some psychological variables could be realized by different neural variables, which is certainly consistent with current neuroscientific practices. Indeed, most cognitive neuroscientific studies pick out the target neural processes by manipulating psychological and behavioral variables and then allow that the neural processes will vary slightly across participants (we're all unique) and even among the same participants over time (see Laumann et al. 2015).[10]

My application of this reasoning simply requires the plausible follow-up suggestion that, for cases of freely willed choices, some of these psychological variables will have the strongest causal invariance relations with decisions and also be ones that, from our own first-person perspective, we would pick out as the sources of our decisions (and fulfill other plausible

[10] Nonetheless, neuroscientists will often be reductionistic in their study of neural mechanisms. So, it is not surprising that some of them think that the neural variables they study have the strongest causal invariance relations with human behavior (e.g., assuming that it is the Readiness Potential, RP, that is the cause of the wrist flex, not my decision to flex or whatever its neural realizers are; Libet 1999). But they are likely wrong to be reductionist in this way, even if neuro-naturalism is true—that is, even if all mental states supervene on neural states. Furthermore, without a Galileo-style theory of consciousness and mental causation, some people may think a dualist metaphysics is the only way to understand the psychological causal interactions (how could a mere meat machine account for Romeo's experience of love?). But of course, non-physical minds or souls do nothing to make sense of these folk psychological explanations. To the extent that those explanations are preserved with a future neuro-naturalistic theory of mind, then we will not become strangers to ourselves. I hasten to add, however, that the details of future neuroscientific discoveries will certainly refine and correct the rough-hewn edges of our self-understanding. And they are likely to indicate that some of us, and some of our decisions, are typically less free than we think (see below).

compatibilist conditions for free will). If the film of my decision to pick B shows that my goal G, in the context of my deliberations, played a crucial causal role, I will not react with surprise or angst. Conversely, if the film showed that some variable I did not know about and would not want to play a crucial role in my decision in fact played a crucial causal role in my decision, my reaction would be quite different. If it showed, for instance, that the neural realizer of a non-conscious priming influence (such as an anchoring effect in the salary offer or the tone of voice of the person offering me the job) had stronger causal invariance relations with my decision, I would see my decision as unfree, since I would not accept this influence were I to know about it, yet I could not control for it since I did not know about it (see Nahmias 2007). Indeed, Greene and Cohen's thought experiment offers a useful way for us to imagine seeing the differences at the neural level that can explain the differences between free and unfree choices (or more and less free actions).

A second application of this interventionist understanding of causal sourcehood uses it to respond the threat allegedly posed by causal determinism. In many cases it is plausible that psychological variables, such as my conscious evaluation of a reason for one option over another, will have a stronger invariance relation with my decision than any of the many causal variables in the past that influence me. If determinism is true, then there is a set of these past causal variables that all together, and in accord with the laws of nature, are sufficient for my decision (this set gets larger and 'wider' the farther back in time we look). Nonetheless, determinism does *not* entail that any of those variables within this set of conditions is the causal source of my decision. Rather, in many cases, events occurring in the 'bottleneck' or nexus of my brain, such as my imagining that job B will be more challenging than A, can be picked out as the causal source of my decision. That deliberative event has a stronger causal invariance relation with my decision than any variables in my distant past. Or being less precise, we might say that the integrated causal activity of my conscious deliberations, as instantiated in the complex nexus of neural activity, is the causal source of my decision. Hence, even if there are causally sufficient conditions in the distant past for my decision, the *causal source* of my decision lies within me, not in any of the causes in my distant past.

This use of interventionist theories of causation can also be used to respond to the most powerful current argument for incompatibilism, the manipulation argument. This argument says that an agent who is manipulated so that he will decide to do B (while satisfying compatibilist conditions) is not free or morally responsible for that decision, but there is no relevant difference between such an agent and one who is causally determined to decide to do B, so the causally determined agent is not free either (Pereboom 2014, Mele 2013). Compatibilists typically respond to this argument by biting the counterintuitive bullet that the manipulated agent is free and responsible. However, the interventionist view of causal sourcehood allows us to uncover a principled distinction between determinism and manipulation. It allows us to see that the manipulated agent's decision has a causal source outside of him—namely, in the intention of the manipulator, since given her knowledge and power, her intention has the strongest invariance relation with his decision. But for the agent in a deterministic universe, there is no causal

variable outside of him (e.g., in his distant past) that is the causal source of his decision. Hence, the compatibilist can use a plausible and generally applicable analysis of causal sourcehood to respond to an incompatibilist argument that aims to conclude that determinism rules out the possibility of our being the causal source of our actions (see Deery and Nahmias, forthcoming, for details).[11]

This way of understanding causal sourcehood in terms of interventionist causal invariance relations and of applying it to defuse potential worries posed by neuro-naturalism or by determinism is technical (though only sketched briefly here). I am not suggesting people are explicitly thinking in these terms when they consider these issues, or when confronted with a neuro-prediction scenario as in Nahmias et al. (2014). After all, I think most people are 'theory-lite'. Nonetheless, it provides a technical way of unpacking our implicit causal cognition such that we can conclude not only that the natural, non-angst-ridden, reaction to neuro-naturalism is, as I've argued, a common reaction, but it is plausibly the correct reaction.

4. Responsibility and Desert

I have not focused here on some issues that are likely the focus of some of the other chapters in this section: moral responsibility, desert, and punishment. I cannot defend here why the view of free will I've outlined secures the types of desert and punishment that some argue are ruled out by determinism (and perhaps neuro-naturalism) (e.g., Pereboom 2014, Pereboom and Caruso this volume). Obviously, a lot turns on how one defines free will and understands the relationship between free will and the relevant notions of desert and punishment. And just as the Copernican theory cannot secure *everything* we once believed, such as our central place in the universe, this view of free will cannot secure some beliefs, such as the misguided ideas that we can be ultimately responsible in some way that might make us deserving of eternal suffering (Strawson 1986) or that we can be uncaused causes in some way that is likely unintelligible.

Nonetheless, my own view is that the neuro-naturalist understanding of free will can support a viable notion of desert that does not depart substantially from most ordinary beliefs and practices regarding moral and legal responsibility (which, unsurprisingly, I suspect are theory-lite as well). Namely, it can support the type of desert that justifies the reactive attitudes—our feelings of gratitude and resentment, pride and guilt—and the related communicative functions of punishment—e.g., holding responsible criminals, who freely do wrong, because they *deserve* to be forced to understand the nature of their crime, to reform to avoid future crimes, to restore the harms they've done as much as possible, and also to express to victims and society the seriousness of those harms (see Nahmias in preparation). While this view does not advocate wrongdoers' suffering for the sake of suffering, as some define retributivist punishment, it does advocate that criminals deserve to suffer to the extent that such suffering is a constitute feature of

[11] The use of causal interventionism might also be used to analyze what it means to say we have the ability to decide otherwise, even in a deterministic universe, and to explain why we experience our future options as open and as causally dependent on what we decide.

these communicative goals of punishing them—for instance, suffering may be a necessary feature (not just a side effect) of the process of coming to understand the harm one has done and feeling and demonstrating appropriate remorse for it.

At the same time, on a naturalistic view of free will, empirical discoveries can inform us about limitations in the relevant capacities and opportunities of humans in general, and also of particular humans. It can thus explain why all of us may be less free and responsible than many tend to assume. And it can explain when and why someone with particular neural deficits (perhaps due to genes or upbringing) thereby lacks the cognitive and emotional capacities to evaluate relevant reasons or control their actions in such a way that it is appropriate to mitigate blame and punishment.

This limited-free-will view may have advantages over pessimism about free will. If many people, in a theory-lite way, associate free will with our capacities for choice and self-control, then when pessimists tell us we have no free will at all, it risks undermining people's belief in those capacities necessary to advocate working hard to improve one's position, to take responsibility for one's failures, to exert willpower in the face of weariness, and to deliberate carefully among alternatives to make good choices—that is, to make personal and moral progress.[12] The limited-free-will view, on the other hand, provides room for such virtues, while it also suggests increased tolerance and compassion for people unfortunate enough to lack sufficient capacities for achieving them. This view can counter an unlimited-free-will view that some people, especially in America, seem to hold, one that suggests people completely deserve everything that happens to them, good or bad, as if they are untethered from the rest of the universe. Realism about the limits of free will, along with a realistic and empirically informed understanding of our capacities, is both more forgiving than an unrealistic theory of unlimited free will and more hopeful and explanatorily fruitful than a pessimism about free will that risks erasing useful distinctions between free and unfree (more or less free) actions.[13]

[12] I think this possibility is plausible without depending on the existing empirical work that has suggested it but that also has various problems in both design and replication (see Schooler et al. 2014; Nadelhoffer et al. this volume).

[13] Portions of this chapter draw on ideas developed in Nahmias (2014), Nahmias and Thompson (2014), and Nahmias, Shepard and Reuter (2014). For helpful comments, I thank Gregg Caruso and Oisin Deery.

References

Campbell, J. 2010. Control Variables and Mental Causation. *Proceedings of the Aristotelian Society*, 110: 15–30.

Caruso, G. 2012. *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will.* Plymouth: Lexington Books.

Deery, O. and Nahmias, E. Forthcoming. Defeating Manipulation Arguments: Interventionist Causation and Compatibilist Sourcehood. *Philosophical Studies.*

Focquaert, F. Glenn, A., and Raine, A. (this volume). Free Will Skepticism, Freedom and Criminal Behavior. In *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, eds. Gregg D. Caruso and Owen Flanagan. New York: Oxford 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.

Gmeindl, L., Yu-Chin Chiu, Michael S. Esterman, Adam S. Greenberg, Susan M. Courtney, Steven Yantis. 2016. Tracking the will to attend: Cortical activity indexes self-generated, voluntary shifts of attention. *Attention, Perception, & Psychophysics*. DOI: 10.3758/s13414-016-1159-7

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

Lagnado, D., T. Gerstenberg, and R. Zultan. 2013. Causal Responsibility and Counterfactuals. *Cognitive Science*, 37: 1036–73.

Laumann T.O., Gordon E.M., Adeyemo B., Snyder A.Z, et al. 2015. Functional system and areal organization of a highly sampled individual human brain. *Neuron* 87(3), 657-70.

Levy, N. 2014. *Consciousness and Moral Responsibility.* Oxford: Oxford University Press.

Libet, B. 1999. Do we have free will? In B. Libet, A. Freeman & K. Sutherland (Eds.), *The Volitional Brain* (47-57). Exeter: Imprint Academic.

List, C. and Menzies, P. 2009. Nonreductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy*, 106: 475–502.

Lombrozo, T. 2010. Causal-Explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions. *Cognitive Psychology*, 61(4): 303–32.

Mele, A. 2013. Manipulation, Moral Responsibility, and Bullet Biting. *Journal of Ethics*, 17(3): 167–84.

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

Mele, A. 2009. *Effective intentions: the power of conscious will*. New York: Oxford University Press.

Monroe, A. & Malle, B. 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, 211-224.

Nagel, T. 1979. *Mortal Questions.* New York: Cambridge University Press.

Nahmias, E., Shepard, J., and Reuter, S. 2014. It's OK if 'My Brain Made Me Do It': People's Intuitions about Free Will and Neuroscientific Prediction. *Cognition* 133(2): 502-513.

Nahmias, E. and Thompson, M. A Naturalistic Vision of Free Will. 2014. In *Current Controversies in Experimental Philosophy*, ed. by E. Machery and E. O'Neill (Routledge), 86-103.

Nahmias, E. Forthcoming. Free Will as a Psychological Accomplishment. In *The Oxford Handbook of Freedom*, ed. by D. Schmitdz & C. Pavel (Oxford University Press).

Nahmias, E. 2014. 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.

Pereboom, D. 2014. *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

Pereboom, D. and Caruso, G. (the volume). Hard-Incompatibilist Existentialism: Neuroscience, Punishment, and Meaning in Life. In *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, eds. Gregg D. Caruso and Owen Flanagan. New York: Oxford University Press.

Rose, D., Buckwalter, W., and Nichols, S. 2015. Neuroscientific Prediction and the Intrusion of Intuitive Metaphysics. *Cognitive Science,* 39 (7).

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

Schooler, J. Nadelhoffer, T., Nahmias, E., and Vohs, K. 2014 Measuring and Manipulating Beliefs about Free Will and Related Concepts: The Good, the Bad, and the Ugly. In *Surrounding Free Will: Philosophy, Psychology, Neuroscience*, ed. by A. Mele (Oxford University Press), 72-94.

Sloman, S. A. 2005. *Causal Models: How People Think about the World and its Alternatives*. New York: Oxford University Press.

Sripada, C. Forthcoming. Free Will and Construction of Options. *Philosophical Studies.*

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

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

Strawson, P. 1962. Freedom and resentment. *Proceedings of the British Academy 48*, 1-25.

Woodward, J. 2015. Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, 91(2): 303–47.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.