# Naturalizing Free Will: Paths and Pitfalls

**Myrto I. Mylopoulos & Hakwan Lau**

**Abstract:** Investigations into the nature of free will have traditionally proceeded on largely theoretical and conceptual grounds. But in recent years, a research program has emerged that aims to develop, refine, and evaluate theories of free will by appeal to methods and data from the natural and social sciences. We call this the *Naturalizing Free Will Program (NFWP)*. This chapter is a critical survey of three of the main implementations of the NFWP: (i) the *Phenomenological Program*, which seeks to get at an accurate description of the phenomenology of free will using the methods of psychology, (ii) the *Intuitionist Program*, which uses the methods of social psychology to systematically investigate folk intuitions surrounding free will, and (iii) the *Cognitive Psychology/Neuroscience Program*, which aims to evaluate theories of free will by appeal to the results and theories of cognitive psychology and neuroscience.

Keywords: free will, naturalism, folk intuitions, phenomenology, experimental philosophy, cognitive neuroscience, cognitive psychology

## §1 Introduction

Theorizing about free will has traditionally proceeded primarily from the armchair. Philosophers working in this area have been content to build their cases and settle their disputes on largely theoretical or conceptual grounds. In recent years, however, a broad research program has emerged that aims to develop, refine, and evaluate theories of free will by appeal to methods and data from the natural and social sciences. We will call this the *Naturalizing Free Will Program (NFWP)*.

The NFWP has at least three distinct subprograms, which will be the focus of this paper. The first attempts to investigate the so-called phenomenology of free will by way of qualitative research methods in psychology, such as the talk-aloud protocol. We call this the *Phenomenological Program.* The second aims to systematically explore folk intuitions pertaining to free will using research methods employed in the social sciences, such as surveys and questionnaires. We call this the *Intuitionist Program*. Finally, researchers have turned to the findings and models of cognitive neuroscience and psychology in order to evaluate theories of free will. We call this the *Cognitive Psychology/Neuroscience Program*.

In this paper, we offer a critical survey of these various attempts to naturalize free will. While we are sympathetic to the overarching aim of the NFWP as a whole, and we take all three of the implementations we will discuss to be of interest and value, we focus here on highlighting some limitations, challenges, and concerns with each of these programs that it will be useful to address moving forward.

## §2 The Phenomenological Program: Free Will and Phenomenology

Theories of free will, it is often urged, ought to answer to how our free will, or lack thereof, subjectively *feels* to us. For example, Nahmias et al. (2004) write that

> [t]heories of free will are more plausible when they capture our intuitions and experiences than when they explain them away. Thus, philosophers generally want their theories of free will to aptly describe the experiences we have when we make choices and feel free and responsible for our actions. If a theory misdescribes our experiences, it may be explaining the wrong phenomenon, and if it suggests that our experiences are illusory, it takes on the burden of explaining this illusion with an error theory (162).

The Phenomenological Program takes on board this prescriptive agenda, and thus aims to arrive at an accurate characterization of the phenomenology of free will, with the further goal of using it to help adjudicate among competing theories. The idea seems to be that, once we have a description of the phenomenology firmly in hand—to the extent that this is possible to acquire—we can credit those theories that align with it, and raise at least *prima facie* doubts about those theories that do not, until a suitable error theory is developed.

In this section, we examine the claim that a theory of free will ought to capture our putative phenomenology. We raise some challenges to the Phenomenological Program on the grounds that (i) there are reasons to doubt that there is indeed a phenomenology of free will, and (ii) even if there is, there are reasons to doubt that it should play any role when deciding among rival theories.[1]

It will be instructive to begin by considering the different ways that theories in general, across domains of inquiry, relate to phenomenology. Some theories take as their main explanandum the structure and central features of our phenomenology. This is certainly true, for example, of the theories developed within the phenomenological tradition by philosophers such as Edmund Husserl (1928/1991), who aimed to elucidate, from the first-person perspective, phenomena like our conscious experience of time and of sensory qualities.

---

[1] One might urge that the role of phenomenology might be construed as helping to pick out the phenomenon to be investigated, thus providing a starting point for theorizing about free will, rather than a basis on which to evaluate such theories. (We are grateful to Eddy Nahmias for pressing this point.) But unless we have an independent grip on the nature of free will in the first place, we cannot determine what is the corresponding phenomenology. So it would seem that we must pick out the concept in a different way to start with.

Evaluating theories of this type must doubtless take into account phenomenological data in a crucial way, since there is a tight relationship between these data and what the theories aim to understand.

By contrast, some theories have distant, or even nonexistent, ties with phenomenology. Consider theories that aim to explain black holes, cell reproduction, or the molecular structure of table salt, to take just a few examples. There is a trivial way in which phenomenology might be thought to be relevant to such theories, i.e., a trivial way in which facts about phenomenology might be necessary for evaluating them: the sensory observations that serve as data for these theories will have phenomenal character associated with them. So, for example, observations of gravitational relations between black holes and other matter will, of course, in virtue of being sensory observations, subjectively seem a certain way. But it would be a stretch to say that such observations, from which the existence of black holes is sometimes inferred, are properly construed as constituting the *phenomenology* of black holes. Phenomenology, apart from in this trivial sense, bears little or no direct relevance to the underlying nature of these phenomena, and so has little or no role to play in evaluating the theories that seek to explain them.

In terms of their relationship to phenomenology, where do theories of free will fall between these two polar extremes? Many would argue that they fit comfortably somewhere in the middle. More specifically, they are viewed as belonging in the camp of theories that do not take as their main explanandum the structure and central features of our phenomenology, but that nonetheless ought to be appropriately *sensitive* to our phenomenology in ways that our theories of black holes need not be. These kinds of theories set out to explain phenomena that have conscious experiences *closely associated* with them. And as such, they should be able to predict or explain features of these experiences. It follows that the relevant class of conscious experiences has, in turn, an evidentiary role to play in adjudicating among competing theories of this type.

On this view, theories of free will are akin to theories of color perception. Arguably, one of the main goals of these theories is to explain how it is that organisms are able to discriminate among the color properties of objects in the environment. But, since color perception often gives rise to a distinctive phenomenology, such theories should also be able to explain and predict the nature of that phenomenology—why it arises in some instances and not others, and why it has the particular features that it does and not others. And our color phenomenology, in turn, arguably takes on a role in evaluating these theories based on how well they predict and explain it. Likewise, it is urged, there are certain characteristic experiences associated with the seeming exercise of one's free will. So a given theory of free will should aim to capture such experiences in the same way, and these experiences should be taken into account when weighing alternative views.

Upon reflection, however, it is far from clear that there is a phenomenology of free will. What could such a phenomenology *be*? Nahmias et al. (2004) write that they will "lump people's experiences of deliberating, making decisions, and feeling free and responsible for their actions together under the

umbrella term 'the phenomenology of free will'" (164). But before adopting this type of strategy, we must take caution in how we treat the various items on this list. It is one thing to claim that people have experiences of deliberating about what to do, of deciding what to do, of consciously intending, and of performing certain actions. It is quite another to claim that they also have experiences of "feeling free" and of feeling "responsible for their actions." The latter claim is much more contentious.

This point becomes clearer once we consider the available candidates for experiences of feeling free and responsible. Free will is characterized in various ways. One of the main ways in which it is characterized is as being compatible or incompatible with determinism. So one primary candidate for an experience of feeling free is an experience of one's action or will as being compatible or incompatible with determinism. But it is doubtful that we have experiences with that content in the absence of a deep internalization of philosophical debates on free will, which is hardly typical. And even if one were to be sufficiently sensitive to philosophical debates on free will, it is still rather mysterious what such experiences would be like. As Richard Holton (2009) puts the worry:

> Sometime [sic] it is said that we have a direct experience of freedom. But if freedom is really understood to be something that is incompatible with determinism, it is hard to know what such an experience would be like. What is it to experience one's action as not causally determined, or oneself as an uncaused cause? I have no idea how that could be the content of an experience.

But perhaps there is another way of understanding the phenomenology free will. Free will has also been traditionally characterized as requiring the ability to do otherwise (see Chisholm, 1964/1997; see also Dennett, 1984; Frankfurt, 1969 for discussion and critique). This characterization seems somewhat more tractable in terms of getting a grip on what it would mean to have a corresponding experience, and in terms of the likelihood of these experiences. When reaching for one's coffee cup with one's left hand, for example, it is tempting to think that one might, if one attends to the possibility, come to have the sense that one could have reached for it with one's right hand instead, or not reached for it at all. And this sense might be aptly labeled the phenomenology of free will.

But here we must be careful about what the relevant content of the sense of being able to do otherwise must be in order for it to count as the phenomenology of free will. It is not just the sense that one could have reached for one's coffee cup with one's right hand; it is the sense that, *keeping the state of the universe fixed prior to one's action*, one could have reached for one's coffee cup with one's right hand. But if this is the kind of phenomenology we are interested in, it is doubtful that people regularly have experiences that exhibit it,

since, once again, it is laden with theoretical complexities that are not typically familiar.

Some will not be moved by the foregoing considerations. They will insist that we *do* have experiences of free will of just the sort that we have been questioning. John Searle (1984), for example, writes in an oft-quoted passage:

> Reflect very carefully on the character of the experiences you have as you engage in normal, everyday human actions... You will sense the possibility of alternative courses of action built into these experiences... that we could be doing something else right here and now, that is all other conditions remaining the same. This, I submit is the source of our own unshakeable conviction of our own free will (95).

Suppose Searle is correct in claiming that experiences of being able to do otherwise, and additional experiences that are aptly characterized as experiences of free will, are more widespread than we have been allowing. What reason would we then have to pay any attention to such experiences when developing theories of free will? Nahmias et al. (2004) argue that the theory and the phenomenology are braided together tightly, since the phenomenology is often used by philosophers as evidence for their theories. And in light of this entanglement between phenomenological claims and theoretical claims, they urge that,

> [i]f possible, then, we need to find out whose descriptions of the experience of free will more accurately reflect pre-philosophical phenomenology. If we find that none does, we need to consider the consequences — for instance, that philosophers should no longer present phenomenology as support for their theory of free will (165).

But this only holds if we establish antecedently that phenomenology *should* have a role to play in theorizing about free will. Otherwise, theorists should stop appealing to it, whether or not their descriptions capture pre-theoretical phenomenology.

This is to challenge the starting assumption of the phenomenological program that theories of free will ought to, in fact, account for phenomenological data. The problem is that theories of free will do not seem to fall in the camp of theories that ought to be sensitive to phenomenology in the way that theories of color perception ought to be, for example. The question is whether theories of free will make any predictions or have anything by way of explanation to say about our putative experiences of free will. Does a libertarian theory, for example, predict that we would have libertarian phenomenology? There is

reason to be doubtful. Libertarian theories hold that free will is incompatible with determinism and that people nonetheless sometimes perform free actions. But this commitment offers no predictions about the character of our experiences. After all, determinism could be false as a result of a quantum event occurring at some point in time, such that it is false that the total state of the universe at any given time combined with the laws of nature entails the total state of the universe at any other time—a common way of understanding the main thesis of determinism (see Mele, 2009). And suppose that this were enough for us to have free will. While this would make some versions of libertarianism true, however, it would clearly not predict anything about our phenomenology, since our phenomenology does not, of course, reflect long-past events at the quantum level.

Perhaps compatibilist theories fare better. On some such views, free will is characterized as the ability to "act according to the determinations of one's will," as David Hume (1748/1993) famously urged in his *Enquiry Concerning Human Understanding*. This might incline one to think that such theories would predict that we experience our actions *as being caused by our intentions, volitions, or desires*. And if so, then perhaps these theories ought to answer to our phenomenology in the way that many suggest. If it never seems to us as though our actions are caused by our intentions, perhaps this is a *prima facie* strike against a Humean compatibilist view. And if it seems to us as though they are, perhaps this is a point in favor.

But once again, there are difficulties. The implicit assumption lurking in the background here is that our conscious experiences infallibly reflect our mental lives, and there is ample reason to doubt this Cartesian assumption based on strong evidence of unconscious mental functioning (e.g., Lau & Passingham, 2006; Lau & Passingham, 2007). So it may be the case that we typically consciously experience our actions as being caused by the "determinations of [our] will" when in fact they are not. Or it may be the case that we typically consciously experience our actions as not being caused in this way when in fact they are.[2] A Humean compatibilist view does not, on its own, predict that one type of conscious experience is more likely than the other.

Indeed, this point applies across the theoretical board. There is reason to be skeptical that our phenomenology is a perfect or even a reliable window into the underlying nature of the self, agency, rationality, and so on. And if it is not, it is risky to assign to it any significant weight when it comes to theorizing about free will, which appeals to such notions. It seems, then, that not only do we have grounds for doubting that there *is* a phenomenology of free will, but if there is, there is reason to doubt that it ought to play a role in evaluating theories of free will.

Still, perhaps these concerns do not mean that the Phenomenological Program, which is still in its infancy, ought to be abandoned. For, if there is

---

[2] Indeed, the psychologist Daniel Wegner has famously argued that this is the case, though we take issue with some of the empirical data he appeals to in support of this conclusion (see §4).

indeed a phenomenology of free will, or phenomenology relevant to free will, perhaps the Phenomenological Program can help us find out. And if so, then even if it cannot help us evaluate theories of free will, it might be valuable for helping us to get a better grip on the subjective character of our experiences.

**§3 The Intuitionist Program: Free Will and Folk Intuitions**

Another camp in the NFWP, the Intuitionist Program, aims to systematically investigate and uncover folk intuitions surrounding free will by way of the methods and techniques employed in the social sciences—most commonly surveys and questionnaires. In a typical set-up, people are presented with hypothetical vignettes and asked to make judgments pertaining to certain features of those vignettes, for example, the moral status of an agent's action. Their answers are taken to reflect pre-theoretical, that is, nonreflective intuitions. For example, in a pioneering study of this sort, Nahmias et al. (2005) presented participants with deterministic scenarios in which an agent performs some salient action. They then asked participants whether the agent in question acted of his or her own free will. They found that participants were significantly more likely than not to judge that the agents featured in these scenarios performed the actions in question "of [their] own free will." The authors conclude that their results "suggest that ordinary people's pre-theoretical intuitions about free will and responsibility do *not* support incompatibilism" (570).

What are we to make of the Intuitionist Program and studies like the one just described? To answer this question, we must get clear on the motivations for the program, otherwise we cannot say whether the program is worthwhile to begin with, nor whether its methods and the data it is producing are serving its aims. One of the central motivations of the Intuitionist Program is that, historically, philosophers have made claims about the intuitions of the folk in support of their own views, without any attempt to back these claims up save for, perhaps, pointing to anecdotal evidence from informal polls conducted in undergraduate philosophy classrooms. For example, Van Inwagen (1993) writes:

> It has seemed obvious to most people who have not been exposed (perhaps 'subjected' would be a better word) to philosophy that free will and determinism are incompatible. It is almost impossible to get beginning students of philosophy to take seriously the idea that there could be such a thing as free will in a deterministic universe. Indeed, people who have not been exposed to philosophy usually understand the word 'determinism' (if they know the word at all) to stand for the thesis that there is no free will. And you might think that the incompatibility of free will and determinism deserves to be obvious— because it *is* obvious (187).

This has led, in the opinion of many, to an unhappy state of affairs in which, as Nahmias et al. (2005) describe it, "philosophers are content to place their *own intuitions* into the mouths of the folk in a way that supports their own position—neglecting to verify whether their intuitions agree with what the majority of non-philosophers *actually* think" (562). So one of the main goals of the Intuitionist Program may be understood as an attempt to determine what folk intuitions *actually* are, such that philosophers' claims pertaining to them may be accurately evaluated.

But while it may be true that philosophers are frequently irresponsible in their appeals to folk intuitions, this observation can only serve as a motivation for the Intuitionist Program if it is also true that folk intuitions somehow *matter* in theorizing about free will. If they are not relevant for theorizing about free will, then the remedy for irresponsible appeals to intuition is not to go out and determine what the folk intuitions really *are*, but for philosophers to simply stop appealing to them altogether. We saw a parallel issue arise in the previous section concerning the Phenomenological Program. There we argued that there is reason to doubt that the phenomenology of free will matters for theorizing about free will. How do folk intuitions fare in this respect?

Murray and Nahmias (2012) make the case that folk intuitions about free will matter because our concept of free will is intimately tied to the conceptual scheme governing our moral practices. They write:

> 'Free will' plays a central role in the conceptual scheme that we use to navigate the normative world via its connections to 'moral responsibility', 'blame', 'autonomy' and related concepts. Theorizing about 'free will' in isolation from the ordinary conception thus risks being an academic exercise about some other, technical concept divorced from people's actual practices of assessing praise, blame, reward, and punishment, and from their understanding of themselves and their place in the world (xx).

There are some worries to be raised here. In everyday morality, when people assess each other's actions or character, or their own actions or character, the concept of free will does not typically play a role. If one is trying to determine whether some agent should be praised or blamed for their conduct, the question of whether or not the agent has free will, or acted out of his or her own free will, does not factor in the deliberation. It rarely makes an appearance in our everyday moral discourse. Instead, other concepts like 'control', 'intention', 'intentional', 'deliberate', 'reason', 'pain', and 'pleasure', seem to be sufficient to "navigate the normative world," as Murray and Nahmias put it. Indeed, many people have never been introduced to the concept of 'free will', let alone the concomitant debates surrounding the notion that have been occupying the attention of philosophers for centuries. And yet their practices of "assessing praise, blame, reward, and punishment" continue without any hindrance. If there

is something essential to these practices that the concept of free will captures, but that other concepts do not, and that makes the concept of 'free will' somehow central to everyday morality, it is important for the proponent of the Intuitionst Program to make clear what that is. As it stands it is not at all so.

Perhaps one will reply that what matters here is not whether the concept of free will itself is regularly appealed to in folk practices governing moral responsibility, but rather that both the folk and philosophers typically take free will to be required for moral responsibility. And what this means is that we must look to folk intuitions about moral responsibility, as well as their practices, to help constrain the concept of free will that figures in philosophical theorizing. On this view, the concept of free will simply refers to whatever type of control is required for moral responsibility. And if so, then when people claim, for example, that unless one acts in a way that is free from constraint, one is not morally responsible for one's actions, they are already tacitly employing the concept of free will, because they are making a claim about the kind of control needed for moral responsibility, and that is just what the concept of free will refers to.

But this line of reasoning faces some difficulties. If the concept of acting with free will has the same extension as, say, the concept of acting in a way that is free from constraint, it does not follow that when one appeals to the latter one is appealing to the former, even tacitly. Analogously, to take a common example, it does not follow that when one believes that the morning star is out, one also believes that the evening star is out, despite the fact that the concept of the morning star and the concept of the evening star have the same extension, that is, the planet Venus. Unless there is reason to think that people in their everyday moral practices and assessments appeal to free will as such, the claim that this concept is deeply engrained in these activities is suspect.

And even if it *were* the case that in appealing to some concept of whatever type of control is required for moral responsibility, one were also appealing to the concept of acting with free will, then the way that the Intuitionist Program is sometimes carried out may need to be revised. For if free will just is whatever type of control is required for moral responsibility, then it may be worthwhile for theorists to eliminate the intermediary step, and only ask people about their intuitions regarding moral responsibility; there is no need for them to concern themselves with how the folk understand free will *per se*.[3]

Another worry for the Intuitionist Program that arises here is whether there really *is* a folk concept or theory of free will that is *antecedent* to philosophical concepts or theories of free will. If not, then it cannot play the role of constraining such theorizing. Consider, by comparison, concepts like 'electron' and 'DNA'. These concepts were not present in folk theories prior to the construction of the relevant scientific theories in physics and biology. The folk, insofar as they have adopted these concepts into their theoretical frameworks, have followed the scientists—not the other way around. Similarly, it may be the case that the

[3] Sometimes theorists do restrict their questions to judgments concerning moral responsibility (e.g., Nichols & Knobe, 2007), but this is not always the case (e.g., Nahmias et al., 2005).

concept of free will was first present in philosophical or, perhaps, theological inquiry, *before* it ever made its way into folk circles. And if so, then looking to the folk to anchor philosophical debates on free will is a backwards or circular enterprise.

A further consideration that puts pressure on the claim that folk intuitions are relevant for theorizing about free will is that folk intuitions are *highly variable*. Across the numerous studies that have attempted to uncover folk intuitions, the results are largely heterogeneous. Some studies turn up largely compatibilist intuitions (e.g., Nahmias, Morris, Nadelhoffer, & Turner, 2006), while others turn up largely incompatibilist intuitions (e.g., Nichols & Knobe, 2007). And all studies turn up some of each.

In addition to *inter*personal variation among folk intuitions, there may also be significant *intra*personal variation—that is, one and the same person may apply different criteria in different cases in order to arrive at judgments pertaining to whether or not an agent has free will or is acting freely. In the closely related domain of moral responsibility, for example, Knobe & Doris (2012) argue that, in fact, at least three factors play a role in determining which criteria people use to form judgments about a given situation: whether the scenario is abstract vs. concrete (e.g., Nichols & Knobe, 2007), the moral valence of the action being performed (e.g., Knobe, 2003), and the relationship between the agent whose action is being judged and the individual making the judgment. Though this hypothesis has not yet been directly tested in the case of free will—and it would be interesting to do so—given the closeness in subject matter, it would be surprising if similar factors were not identified with respect to folk intuitions in this area as well.

The foregoing suggests that the idea that there exists *the* folk concept or theory of free will, which many seem to assume, is very likely a fantasy. Instead, it would seem that there are *many* such concepts or theories.[4] But if so, then there is reason to doubt claims to the effect that a particular theory of free will comports with *the* folk notion of free will—there does not seem to be such a thing. Instead, individual theories of free will should be taken, to the extent that they do, to account for a *slice* of folk intuitions, on the understanding that these are but slivers of the pie. And if so, then it is unclear that folk intuitions can play the role that they have been assigned of supplying even *prima facie* evidence for or against theories of free will insofar as they align with them. If a theory comports with some folk intuitions but not others, then this says little or nothing about the merits of that particular theory as against other theories, since, given that folk intuitions are significantly variable, serious competing theories will also capture some folk intuitions.

---

[4] Recent work by Monroe and Malle (2010) suggests that there might be a majority concept of free will as "a choice that fulfills one's desires and is free from internal or external constraints" (211). But it is not clear why a theory that reflects a majority intuition should be preferred over one that reflects a significantly held, but minority intuition. So this does not help secure an adjudicating role for intuitions in evaluating theories of free will.

One might object here that much of the seeming variance in intuitions can be explained away. In the face of conflicting data on folk intuitions, theorists often put forward "error theories" to account for the discrepancies, running further studies in attempts to corroborate them. For example, responding to results suggesting that people tend towards incompatiblist intuitions, Murray and Nahmias (2012) propose that some people mistakenly take determinism to entail what they call "bypassing". This involves viewing determinism as entailing that "rational deliberation, conscious consideration of beliefs and desires, formation of higher-order volitions, planning" (xx) and other capacities that compatibilists identify as essential for free will play no role in producing an action. And so, according to Murray and Nahmias, when asked if an agent is free or morally responsible in a deterministic scenario, some individuals, assuming determinism to entail bypassing, answer 'no', when in fact, were they to understand that this entailment does not hold, they would answer 'yes'. Murray and Nahmias thereby claim that some individuals express *apparent*, but not *genuine* incompatibilist intuitions, since irrelevant factors, in this case a misunderstanding of the technical concept of determinism, are responsible for the judgments in question.

But if "genuine" intuitions are those that arise from a solid understanding of the theoretical terrain, and remaining intuitions are merely apparent and not to be used in theorizing, far from saving the import of folk intuitions, this strategy suggests that we ought to abandon them altogether. After all, the folk typically do *not* have a robust grasp of the relevant theoretical issues, involving as they so often do technical concepts like 'determinism', the 'ability to do otherwise', 'reasons responsiveness', 'second-order desires', and so on. They may have a reasonable grasp of non-technical concepts like 'moral responsibility', but what theorists are typically interested in is how such concepts *relate* to the technical ones that they propose. And if so, it is not clear that the folk can help.

Perhaps, though, this does not entail setting aside intuitions altogether. Rather, one might conclude from the foregoing that we ought still to look to those whom Mele (2006) has labeled "reflective agnostics" (191) for their intuitions. Reflective agnostics are people who have thought carefully about the relevant debates, but have yet to make up their minds about what the right thing is to say. While reflective agnostics will certainly be better off than the folk in having a clear understanding of the technical machinery underlying debates in free will, there is at least one significant reason to think that their intuitions cannot play the role of helping to resolve theoretical disputes. The reason is that their intuitions are arguably the *result* of the theories that they are entertaining, whether tacitly or not. Certainly, intuitions have the subjective *appearance* of immediate, unreflective judgments. But, given their status as judgments, they cannot actually be free floating and detached. Rather, they must stand in inferential relations with a whole network of other mental states within one's mental economy. Indeed, this is a premise of the Intuitionist Program, since collecting people's intuitions is supposed to reveal the "contours" of their folk theories. In the case of the reflective agnostic, whatever "contours" are revealed by their intuitions will belong to whatever theories they have been puzzling over. Their

intuitions are the deliverances of those theories, and so they cannot serve as evidence for them.

## §4 The Cognitive Psychology/Neuroscience Program

We turn now to a third approach in the NFWP, which seeks to draw on the methods and results of neuroscience and cognitive psychology to answer key questions that arise in theorizing about free will. Just like the Phenomenological Program and Intuitionist Program we have been discussing, this program faces its own set of challenges and obstacles, but we believe it has particular promise.

Neuroscience and psychology, of course, has their limitations when it comes to helping us settle questions surrounding free will. Roskies (2006) argues that neuroscience is not informative, for example, with respect to the question of whether the universe is deterministic. The problem is that neuroscience explains phenomena at the level of the brain, i.e., at the level of neurons, synapses, and action potentials. But, as Roskies points out, apparent determinism at this level of explanation is compatible with actual *in*determinism at lower levels, which explain the world in terms of atomic or subatomic particles, and vice versa. As such, the question of determinism will ultimately be settled by physical theories that aim to give a full account of the fundamental level of reality, not by neuroscientific theories that aim to capture what is happening at a level higher up.

Where neuroscience and psychology *can* be of service, however, is in providing theoretical models of deliberation, decision-making, action control, and consciousness—among other phenomena—and in collecting empirical data to help evaluate these models. After all, these psychological functions and features often play essential roles in accounts of free will—especially compatibilist accounts. So having a clear understanding of how they operate is paramount for adjudicating between rival theories. This is one of the main reasons that we find the Cognitive Neuroscience Program to be especially valuable.

In the past few decades, researchers have already been undertaking the significant project of connecting advances in neuroscience and psychology with issues pertaining to free will. At the center of much of this activity is the important question of what role consciousness plays in free will. Many theorists suppose that freely performed actions require some contribution from consciousness. For example, the psychologist William Banks (2006) wrote that "[f]ree will seems pointless if it is not conscious free will. We are not interested in unconscious freedom of the will, if there is such a thing…" (236). And the philosopher John Searle (2010) writes, "only for the conscious agent can there be such a thing as freedom of the will" (129).

Traditionally, this assumed link between consciousness and free will has not been the focus of much attention, as it was thought to be unproblematic to assume that consciousness plays some role in deliberation and action, and that this matters for free will. But some of the work coming out of the Cognitive Psychology/Neuroscience Program purports to challenge the former assumption. Perhaps most influentially, the neuroscientist Benjamin Libet (1983) and his

colleagues claimed that our basic actions, e.g., a flexing of the wrist, are initiated by a neural event (the Readiness Potential or RP) that takes place approximately 300 ms *prior* to a conscious decision to act (see also Kornhuber & Deecke,1965). As Libet (1985) himself put it, "the brain 'decides' to initiate or, at least, to prepare to initiate the act before there is any reportable subjective awareness that such a decision has taken place" (536).

An equally bold claim has famously been put forward by the psychologist Daniel Wegner (2002), who argued that our experience of consciously willing our actions is illusory—the product of a psychological mechanism of causal reasoning that we apply to ourselves, and which falsely characterizes our conscious intentions as causing our actions. Many have viewed the possibility that consciousness, specifically as a property of mental states, does not play the role that it seems in producing our actions as a threat to free will (see Mele, 2013 for a useful discussion of whether such anxieties are warranted).

Though Libet and Wegner's work is doubtless important, there has been a myopic focus on it in the literature, despite several limitations that it faces. To start with Libet, there have been a number of influential critiques of his work, which we will mention only briefly here. One source of disagreement concerns whether participants' introspective timing reports are reliable (e.g., Banks & Isham, 2009; Lau, Rogers, & Passingham, 2007). Others have taken issue with the claim that the RP is indeed the neural signature of action initiation (e.g., Schurger, Sitt, & Dehaene, 2012; Trevena & Miller, 2010), or they find fault with particular features of Libet's methodology (e.g., Gomes, 1998).

In addition, as others have pointed out, Libet's results do not straightforwardly generalize to the role of consciousness in producing intentional actions more broadly. Libet et al. (1983) asked participants to "let the urge to act appear on its own at any time without any preplanning or concentration on when to act" (625). These instructions were supposed to ensure that participants performed actions that were "freely capricious in origin" (625). But in attempting to design the perfect "spontaneous" act, Libet and colleagues rendered their results problematically narrow in scope and application. The vast majority of our actions are *not* spontaneous basic actions that are performed in the absence of any plan. They are typically nonbasic actions, i.e., those that one cannot perform without doing something else first (e.g., crossing the street) and they are typically performed in the service of some antecedently formed plan (e.g., going to the park), however simple. It is unclear whether Libet's work does anything to illuminate the nature of *these* actions and their relation to consciousness, but it would seem that it is these actions that are relevant to the question of whether or not we have free will. (See also Mele, 2009a for a sweeping and careful critique of Libet's interpretation of his results.)

While Libet's results are, among other things, too narrow in scope when it comes to their implications, they are at least somewhat robust; the main result, of the RP preceding the time at which participants report having decided to act, has been replicated and refined (e.g., Haggard & Eimer, 1999; Lau, Rogers, Haggard, & Passingham, 2004). Wegner's main studies, however, have yet to be replicated, though experiments using similar paradigms have been carried out

and arrived at compatible findings (e.g., Aarts, Custers, and Wegner, 2005; Sato, 2009; Wenke, Fleming, and Haggard, 2010). Indeed, there are methodological issues with Wegner's most widely cited studies that are not frequently addressed or taken into account. One of these studies is Wegner & Wheatley's (1999) 'I Spy' study. In this study, participants were paired with confederates from whom they sat across, with a square board in between them that was mounted on top of a computer mouse. Both the participant and the confederate were asked to place their fingertips on the board so as to move the mouse together, simulating a 'ouija board' set up. They were asked to move the mouse in "slow sweeping circles," which would move a cursor on a computer screen that they could both see. On the screen was a photo showing a number of small objects (e.g., a car, a plastic dinosaur).

Participants and confederates were asked to stop moving the mouse every 30 seconds, after which they would rate how much they intended to make the stop. They did so by recording marks on a line that had one endpoint indicating 'I allowed the stop to happen' and another endpoint indicating 'I intended to make the stop'. The participants' marks on the line were afterwards converted to percentages between 0 – 100 by the experimenters.

In more detail, after each 30-second interval, there was a 10-second interval during which the participants and confederate were supposed to make a stop. During this "stop" interval, the participants would hear music and a single word over their headphones, sometimes naming some object on the screen. They were told that they were hearing different words than the confederate, and that the words were meant to serve as "mild distractions." In fact, on some trials, the confederate was hearing instructions to move to a particular object on the screen at a particular time. The timing was such that the participant would hear the word corresponding to the object the confederate stopped on either 30s, 5s, 1s before, or 1s after the confederate stopped on the object. On the rest of the trials, the confederate let the participant make the stops. In these cases, the participants heard a word two seconds into the 10-second "stop" interval. The word corresponded to an object on the screen for roughly half the trials only, to make it credible that they were merely meant to be "distractions."

The participants rated the "forced" stops, that is, those stops made by the confederate, at around 52% on the scale between 'I allowed the stop to happen' and 'I intended to make the stop'. Moreover, the degree to which they rated the stops as intended increased the closer the priming word occurred to the stop, with the average rating at around 44% when it occurred 30-seconds before the stop, and climbing up to between 55 – 60% as it approached 5-seconds and 1-second before, then dropping down again to approximately 45% when it occurred 1-second after the stop. From this, the authors conclude that "… there was a tendency overall for participants to perceive the forced stops as intended" (489).

But this interpretation of the results is not warranted. The participants on average barely rated the intentional nature of the relevant stop as more than halfway between allowing it to happen and its being intended. If they believed that they intended the stop, one would expect them to rate it at or near 100%, on

the side of the line that is explicitly labeled 'I intended to make the stop'. Given that they did not, it is doubtful that they believed that they intended the stop.

Another issue with the 'I spy' study is that a second agent is potentially contributing to the action in question, resulting in a highly ambiguous context. Indeed, the participants seemed sensitive to the ambiguity of the situation, given their mid-range ratings of the intentional character of the stop. One explanation of these ratings is that participants neither judged that they allowed the stop to happen, nor judged that they intended it to happen. This might be the case, for example, if they had no intention of stopping the cursor on a particular item, but felt that they still contributed to the stop by moving the cursor jointly with the experimenter. A follow-up experiment could test for this alternative possibility by giving participants the option of answering that they neither intended the stop nor allowed it to happen. As it stands, these two options do not exhaust the possibilities.

Another widely cited and endorsed study is the 'helping hands' study conducted by Wegner, Sparrow, and Winerman (2004). In this study, participants watched themselves in the mirror while another subject—a "hand helper"—stood behind them and extended their hands forward on either side of the participant. From the participant's point of view, another person's arms and hands were located where their own arms and hands would normally be.

The "hand helpers" heard a sequence of instructions over headphones, such as "wave hello with your right hand," and "give the OK sign with both hands." Participants were told that they may or may not hear instructions over their headphones, and that if they did hear instructions, they may or may not relate to the actions of the hand helper. In one condition (preview condition), participants heard the instructions at the same time that the hand helper followed them. In the other condition (no preview condition), participants heard nothing through the headphones.

Afterwards, participants were asked to rate their experiences, based on different questions, on a 7-point scale from 1 ("not at all") to 7 ("very much"). The key questions, for our purposes were, "How much control did you feel that you had over the arms' movements?" and "To what degree did you feel you were consciously willing the arms to move?" These questions were thought to measure the participants' sense of control regarding the movements of the hands. The responses to these questions were correlated, so Wegner et al. (2004) took the mean of these responses as "an index of vicarious agency." The authors report that "[i]n line with our hypothesis, the participants receiving previews expressed an enhanced feeling that they were able to control and will the arms' movements" (841). They conclude this on the grounds that "mean vicarious control ratings" were significantly greater with previews ($M = 3.00$, $SD = 1.09$) than without ($M = 2.05$, $SD = 1.61$).

There are problems with this study as well, however. For one, it is not clear that what is being probed is a sense of control over the arm movements. Instead, the questions asked could be probing 'as if' judgments. The participants plainly did not believe that they actually were controlling the movements of the experimenter. They did not believe that the arms of the experimenter were their

arms, nor did they believe that any of their mental states could somehow exert control over another agent's arms—it would be incredible to suppose otherwise. As a result, in answering question 2, they may have reasoned as follows: "If I *were* controlling the arms, I would have an idea of what movement they were about to perform before performing them. To the extent that I have an idea of what movement the arms are about to perform before they perform it, I judge that it is *as if* I am in control of the arm movements. But, of course, I am not actually in control of the arm movements, nor do I experience myself as such." (A similar thought process might have accompanied their answers to question 3.)

Second, the ratings given by the participants were very low even in the preview condition. They were exactly at the 3-point mark, which is less than halfway up the 7-point scale being used. This suggests that participants did not actually feel that they controlled the hand movements, nor did they feel that they were consciously willing the hands to move. This is worth stressing, as some theorists have not been sensitive to the low value of the ratings in discussing their interpretation of the results, thereby exaggerating the significance of the results that were found (e.g., Synofzik, Vosgerau, & Newen, 2008, pp.226)

There are reasons to doubt, therefore, that the results from these two frequently cited studies, the 'I Spy' study and the 'helping hands' study, should be taken as evidence that participants in the experiments were caused to erroneously experience willing actions that they did not, in fact, will, as many in the literature have claimed (e.g., Prinz, 2012, pp.191). And even if they could, it would be a further step to argue that this is the case more generally. Showing that people in experimental settings sometimes have illusory conscious willings is not, of course, sufficient to establish that conscious willings are always or even typically illusory. And if so, then such findings cannot support the view that our conscious intentions do not cause our actions.[5]

Still, despite these shortcomings, both Libet and Wegner, and those who have followed in their footsteps, aim to answer a very pressing question: what is the role of consciousness in producing our actions?[6] If cognitive neuroscience and psychology can deliver an answer to this question, with suitably refined methods (see, e.g., Lau & Passingham, 2007; van Gaal et al., 2010) then combined with an answer to the question of whether consciousness is required for free will in the first place, this would yield a concrete, straightforward development in the study of free will. This work is thus illustrative of why we find

---

[5] See also Shepherd (2013), Malle (2006), and Nahmias (2005) for similar critiques of these studies.

[6] It is important to keep in mind that this is a separate question from the role of phenomenology in adjudicating between theories of free will, which we addressed in §2. It might be true that phenomenology plays no such role, while remaining true that in order to act freely, one's actions must be caused by one's conscious intentions, decisions, and so on. One is a question about how it is that we are conscious *of* our own free will, and the other is a question about what role being in conscious states plays in securing free will.

the Cognitive Psychology/Neuroscience Program exceptionally promising as an avenue of research in the NFWP.

## §5 Conclusion

In this chapter, we have critically surveyed three of the main approaches to naturalizing free will. As mentioned, we take all three of these to be of interest and of importance. We do hope, however, that as the NFWP proceeds and expands in the years to come, some of the challenges and issues we have raised here will be worth taking into account.

## Acknowledgements

We are grateful to Al Mele, Eddy Nahmias, and Joshua Shepherd for helpful comments on earlier drafts of this chapter.

This chapter was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this chapter are our own and do not necessarily reflect the views of the John Templeton Foundation.

**Sources**


Aarts, H., Custers, R., & Wegner, D. M. (2005). On the inference of personal authorship: enhancing experienced agency by priming effect information. *Consciousness and Cognition, 14*(3), 439-458.

Banks, W. P. (2006). Does consciousness cause misbehavior? In S. Pockett, W. P. Banks & S. Gallagher (Eds.), *Does consciousness cause behavior?* (pp. 235 - 256). Cambridge, MA: MIT Press.

Chisholm, R. (1997). Human freedom and the self. In D. Pereboom (Ed.), *Free will* (pp. 24 - 25). Indianapolis, Indiana: Hackett Publishing Company, Inc. Original published in 1964.

Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility *Journal of Philosophy, 66*(23), 829 - 839.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*(1), 128-133.

Holton, R. (2009). Determinism, self-efficacy, and the phenomenology of free will. *Enquiry, 52*(4), 412 - 428.

Hume, D. (1993). *An enquiry concerning human understanding*. Indianapolis, Indiana: Hackett Publishing Company, Inc. Original published in 1748.

Husserl, E. (1991). On the phenomenology of the consciousness of internal time (1893 - 1917)(J. B. Brough, Trans.). Translated by John Barnett Brough. Dordrecht, Netherlands: Kluwer Academic Publishers. Original published in 1928.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190 - 193.

Knobe, J., & Doris, J. M. (2012). Strawsonian variations: Folk morality and the search for a unified theory. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 321 - 354). Oxford, UK: Oxford University Press.

Kornhuber, H. H., & Deecke, L. (1965). Hirnpotentialänderungen bei willkürbewegungen und passiven bewegungen des menschen: Bereitschaftspotential und reafferente potentiale. (Changes in brain potentials with willful and passive movements in humans: the readiness potential and reafferent potentials.). *Pflügers Archive*(284), 1 - 17.

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *PNAS*, *103* (December (49), 18763–18768.

Lau, H. C. & Passingham, R. E. (2007). Unconscious activation of the cognitive control system in the human prefrontal cortex. *Journal of Neuroscience, 27*(21), 5805 - 5811.

Lau, H. C., Rogers, R. D., Haggard, P., & Passingham, R. E. (2004). Attention to intention. *Science, 303*, 1208 - 1210.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential).

The unconscious initiation of a freely voluntary act. *Brain, 106 (Pt 3)*, 623-642.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences*, 8, 529 - 566.

Malle, B. F. (2006). Of windmills and straw men: Folk assumptions of mind and action. In S. Pockett, W. P. Banks & S. Gallagher (Eds.), *Does consciousness cause behavior?* (pp. 207 - 231). Cambridge, MA: MIT Press.

Mele, A. R. (2006). *Free will and luck*. Oxford, UK: Oxford University Press.

Mele, A. R. (2009a). *Effective intentions: The power of conscious will*. Oxford, UK: Oxford University Press.

Mele, A. R. (2009b). Free will. *Encyclopedia of Consciousness*, *1*, 265 - 277.

Mele, A. (2013). Unconscious decisions and free will. *Philosophical Psychology*, *26*(6), 777 - 789.

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*, 211 - 224.

Nahmias, E. (2005). Agency, authorship, and illusion. *Consciousness and Cognition, 14*, 771 - 785.

Murray, D., & Nahmias, E. (2012). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2004). The phenomenology of free will. *Journal of Consciousness Studies, 11*(7 - 8), 162 - 179.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561 - 584.

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.

Prinz, J. (2012). *The conscious brain*. Oxford, UK: Oxford University Press.

Roskies, A. (2006). Neuroscientific challenges to free will and moral responsibility *Trends in Cognitive Science, 10*(9), 419 - 423.

Sato, A. (2009). Both motor prediction and conceptual congruency between preview and action-effect contribute to explicit judgment of agency. *Cognition*, 110(1), 74 - 83.

Searle, J. (1984). *Minds, brains, and science*. Cambridge, MA: Harvard University Press.

Searle, J. (2010). Consciousness and the problem of free will. In A. R. M. Roy F. Baumeister, Kathleen D. Vohs (Ed.), *Free will and consciousness: How might they work?* New York: Oxford University Press.

Shepherd, J. (2013). The apparent illusion of conscious deciding. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action, 16*(1), 18 - 30.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition, 17*(1), 219 - 239.

Van Gaal, S., Ridderinkhof, K. R., Scholte, H. S., Lamme, V. A. F. Unconscious activation of the prefrontal no-go network. *Journal of Neuroscience, 30*(11), 4143 - 4150.

Van Inwagen, P. (1993). *Metaphysics* (1st ed.). Boulder, CO: Westview Press.

Wegner, D. (2002). *The illusion of conscious will*. Cambridge, MA: Bradford Books.

Wegner, D. M., Sparrow, B., & Winerman, L. (2004). Vicarious agency: experiencing control over the movements of others. *Journal of Personality and Social Psychology, 86*(6), 838-848.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation. Sources of the experience of will. *American Psychologist, 54*(7), 480 - 492.

Wenke, D., Fleming, S. M., & Haggard, P. (2010). Subliminal priming of actions influences sense of control over effects of action. *Cognition, 115*(1), 26-38.