Philosophy Compass 5/2 (2010): 199–212, 10.1111/j.1747-9991.2009.00273.x

Experimental Philosophy and Free Will

Tamler Sommers*
University of Houston

Abstract

This paper develops a sympathetic critique of recent experimental work on free will and moral responsibility. Section 1 offers a brief defense of the relevance of experimental philosophy to the free will debate. Section 2 reviews a series of articles in the experimental literature that probe intuitions about the ''compatibility question''—whether we can be free and morally responsible if determinism is true. Section 3 argues that these studies have produced valuable insights on the factors that influence our judgments on the compatibility question, but that their general approach suffers from significant practical and philosophical difficulties. Section 4 reviews experimental work addressing other aspects of the free will/moral responsibility debate, and section 5 concludes with a discussion of avenues for further research.

1. Introduction

The precise meaning of 'experimental philosophy' is a matter of controversy even among its practitioners, but all agree that it refers to a movement which employs the methods of empirical science to shed light on philosophical debates. Most commonly, experimental philosophers attempt to probe ordinary intuitions about a particular case or question in hopes of learning about the psychological processes that underlie these intuitions (Knobe 2007).[1] The unifying conviction behind the movement is that many deep philosophical problems 'can only be properly addressed by immersing oneself in the messy, contingent, highly variable truths about how human beings really are'.[2] And the problem that has received the most attention from experimental philosophers thus far is the problem of free will.

My aim in this paper was to offer a sympathetic critique of recent experimental work on free will and moral responsibility. I begin in section 2 by offering a brief defense of the relevance of experimental philosophy to the free will debate. In section 3, I review a series of articles in the experimental literature that probe intuitions about the 'compatibility question' – whether we can be free and morally responsible if determinism is true. In section 4, I argue that although this work has produced valuable insights into the factors that influence our judgments on the compatibility question, the general approach of these studies suffers from significant practical and philosophical difficulties. Section 5 reviews experimental work addressing other aspects of the free will/moral responsibility debate, and section 6 concludes with a discussion of avenues for further research.

2. Why experimental philosophy is relevant

Philosophers from the 'armchair' tradition who work on free will tend to be supportive of experimental approaches, far more so than in other areas of philosophy. Still, every so

often one hears a version of the objection: 'I don't care about people's intuitions about free will and moral responsibility, I'm interested in the *truth* about free will and moral responsibility. Experimental philosophy doesn't tell me anything about that!' At first glance, this kind of objection may seem to have a sensible ring to it. In debates over, say, group selection in evolutionary theory, we do not proceed by examining folk intuitions about how Darwinian natural selection might work. Why shouldn't level-headed, no-nonsense philosophers regard free will and moral responsibility the same way?

The answer is simple: unlike evolutionary biologists, philosophers have thus far investigated the nature of their topic through an appeal to the intuitions of their audience. Arguments for incompatibilism about free will and moral responsibility rely on some version of the principle of alternate possibilities (PAP) or, more commonly, a 'transfer of powerlessness' or 'transfer of non-responsibility' principle.[3] Incompatibilists often appeal directly to the intuitive plausibility of these principles (e.g. Van Inwagen 1983; Strawson 1986)[4] or they describe specific cases in which an agent is not free or morally responsible and then argue that there are no relevant differences between those cases and *all* instances of fully determined human behavior (e.g. Pereboom 2001). [Correction added after online publication 10 February 2010: Sentence changed.] For the latter 'generalization strategies' (as R.J. Wallace has called them) to work, the reader must share the starting intuitions that the agents in the original cases are not free and morally responsible, and agree that there are no intuitively plausible differences between the cases and determined behavior in general.[5]

Compatibilists are no less reliant on appeals to intuitions, both to develop counterexamples to incompatibilist principles[6] and to develop sufficient conditions for free and responsible behavior. The structure of Wolf's (1987) argument for the 'sane deep self-view' of moral responsibility provides a nice illustration of the role of intuitions in compatibilist theorizing. Wolf describes the case of JoJo, the son of a brutal dictator, who has been trained from early childhood to value arbitrary expressions of cruelty, such as executing or torturing his subjects on the basis of mere whim. JoJo understandably sees his father as a role model and acquires a character that values cruel behavior as well – thus meeting the conditions for moral responsibility established by the 'deep self-view' of Frankfurt and Watson.[7] According to Wolf, however, 'in light of JoJo's heritage and upbringing, it is dubious at best that he should be regarded as responsible for what he does' (pp. 379–380). This leads Wolf to add a condition to the deep self-view: the condition of sanity, which includes a capacity to understand the difference between morally right and wrong action. Wolf's argument has two premises that require intuitive agreement. First, we must share the (perhaps controversial) intuition that because of his upbringing, JoJo is not morally responsible for his cruel behavior. Second, we must agree that the addition of the 'sanity requirement' provides the deep self-view with sufficient conditions for moral responsibility. (And to argue against Wolf's 'sane deep self' theory one must devise an intuitively plausible counterexample. The cycle continues.)

The bottom line is that contemporary theories of free will and moral responsibility essentially involve appeals to intuitions about key principles and cases. So, if we are interested in the truth about free will and moral responsibility, we must maintain a lively interest in the intuitions and beliefs of others. Experimental methods shed light on: (a) the psychological mechanisms underlying these intuitions, and (b) the degree to which the intuitions that drive philosophical theorizing are shared by members of a larger community. Experimental philosophy, then, can help us understand the nature of free will and moral responsibility in ways that complement more traditional 'armchair' analysis.[8]

3. The Central Dialectic: Probing Intuitions on the Compatibility Question

For the remainder of this essay, I will assume the relevance of experimental inquiry to the free will debate in general and focus on specific trends that have characterized recent experimental work on the topic. Since the beginning of the modern period, philosophers have focused obsessively on the question of whether determinism and free will are compatible, so it is not surprising that experimentalists have followed suit. Beginning with Nahmias, Morris, Nadelhoffer, and Turner's (NMNT) seminal studies, philosophers have attempted to directly probe folk intuitions on the compatibility question, and subsequent experiments have been developed to challenge, support, reinterpret, and shed new light on their results. Nahmias et al. (2005, 2006) presented subjects with a series of vignettes in which an agent performs a moral or immoral action in a determined world. The first study used a Laplacian description of determinism: subjects were told that scientists in the next century have discovered the laws of nature and developed a supercomputer that can predict future events with 100% accuracy. The supercomputer predicts that a man named Jeremy will rob a bank at 6:00 PM on January 26, and, as always, the supercomputer is right. Subjects were then asked whether Jeremy acted of his own free will and whether he is morally blameworthy for robbing the bank. Somewhat surprisingly, a large majority of the subjects gave compatibilist answers to both questions – 76% judged that Jeremy acted of his own free will and 83% responded that he was morally blameworthy for robbing the bank. Nahmias and colleagues received complementary results for the praiseworthy and morally neutral scenarios as well. To make sure that the determinism in the scenarios was sufficiently salient, the authors developed two more ways of describing determinism, one that involved a universe that is recreated over and over again with the same initial conditions and laws producing the same events each time, and another that appealed to genetic and environmental influences. Again, the authors received largely compatibilist responses from their subjects. According to NMNT, these results cast doubt on the claims of philosophers, such as Robert Kane and Galen Strawson, that the folk are 'natural incompatibilists' and should, at the very least, shift the burden of proof towards the incompatibilist side.[9]

Nichols and Knobe (2007) adopt the same basic approach in their article 'Moral Responsibility and Determinism: the Cognitive Science of Folk Intuitions'. They note the asymmetry between incompatibilist claims about folk intuitions and the experimental results in NMNT but suggest that they can be reconciled when we consider the role of affect in generating folk judgments about moral responsibility. In philosophy seminars, when considering the compatibility question abstractly, Nichols and Knobe suggest, we might have incompatibilist intuitions. However, in concrete cases that trigger emotional responses, our intuitions become more compatibilist. To test this hypothesis, Nichols and Knobe describe two universes to their subjects, a deterministic universe (Universe A) in which everything including human decision making is completely caused by events tracing all the way back to the beginning of the universe, and the other (Universe B) where everything *with the exception* of human decisions is completely caused by past events. The key difference, according to the scenarios:

> is that in Universe A every decision is completely caused by what happened before the decision – given the past, each decision *has to happen* the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision *does* ***not*** *have to happen* the way that it does (p. 669).

The subjects are first asked which universe more resembles our own and over 90% of the subjects respond that it is Universe B, the indeterministic universe. The subjects are then

split into abstract and concrete conditions. In the abstract condition, subjects are asked: 'in Universe A, is it possible for a person to be fully morally responsible for their actions?' Here, a large majority (86%) of the subjects answer 'no', the incompatibilist response. In the concrete condition, subjects are told that in the deterministic universe, a man named Bill burns down his house, killing his wife and three children, in order to be with his secretary. They are then asked if Bill is fully morally responsible for his behavior. In this condition, 72% of the subjects gave the compatibilist response, judging that Bill is morally responsible for his horrifying crime. Nichols and Knobe then conducted a follow-up study which suggested that subjects were far more likely to give incompatibilist responses in low-affect concrete cases than high-affect cases (64% to 23% respectively). On the basis of this result, the authors tentatively conclude that a 'performance error' model best explains their results. The emotions triggered in high-affect concrete cases hampers the subjects' ability to correctly apply their incompatibilist intuitions. While noting that the results do not offer anything like decisive support for incompatibilism, the Nichols and Knobe argue that the study could provide a debunking explanation for the compatibilist intuitions generated in the NMNT cases (all of which are concrete, although not as affect laden).

Eddy Nahmias and a new set of collaborators find fault with Nichols and Knobe's description of determinism, specifically the claim that 'everything *has to happen* the way it does'. In their paper 'Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions' (Nahmias et al. 2007), NCK suggest that this description may suggest to subjects that desires and conscious deliberation are *bypassed*, that is, causally irrelevant to our resulting behavior. As determinism itself does not entail that our desires and deliberation are not part of the causal process leading to behavior, the incompatibilist intuitions generated in the abstract condition may be a result of the subjects confusing determinism with *fatalism*. If this is the case, the concrete condition in Nichols and Knobe's study would still generate a performance error, but not one that occurs in judgments about a deterministic world. Rather, the error would apply to judgments about a fatalistic world where desires, goals, and reasoned deliberation are causally irrelevant to the resulting behavior.

To support this interpretation, NCK developed two kinds of deterministic scenarios – one in which the bypassing threat is present and one in which it is absent. The authors operate under the assumption that when the decision-making process is described mechanistically, as neural processes and chemical reactions in the agent's brain, subjects tend to think that beliefs, desires, and reasoning are causally impotent. This assumption provides the key manipulation for the deterministic scenarios. In one condition, the agents' decision making is described 'in terms of neuroscientific, mechanistic processes (Neuro scenarios)'; in the other, decision making is described 'in terms of psychological, intentional processes (Psych scenarios)'. NCK made three central predictions:

1. That most people will judge that determinism is *not* threatening to FW and MR if determinism is described in *non*mechanistic (psychological) terms.
2. That significantly more people *will* judge determinism to be threatening to FW and MR if determinism is described in mechanistic (neuroscientific) terms.
3. That people will significantly increase their judgments of FW and MR (in both conditions) in response to descriptions of specific agents who perform bad acts in comparison with agents and actions described in abstract ways (p. 221.)

The results appear to offer significant support for their predictions. Subjects tended to give compatibilist responses in the abstract/psych condition but incompatibilist responses

in the abstract/neuro condition. Neuroscientific descriptions of decision making seem to be more of a threat to responsibility than psychological ones.

The problem, however, is that just as Nahmias and colleagues had grounds to object to 'had to happen' language, the incompatibilist may protest that the psychological descriptions do not make the determinism sufficiently salient. Consider the final paragraph of the description:

> So, if these psychologists are right, then once specific earlier events have occurred in a person's life, these events will definitely cause specific later events to occur. For instance, once specific thoughts, desires, and plans occur in the person's mind, they will definitely cause the person to make the specific decision he or she makes (p. 224).

As NCK here focus on perhaps the least threatening aspect of determinism (that our thoughts and desires determine our actions), the description may not provide enough emphasis on the historical aspect of determinism, the notion that these thoughts, desires, and plans are determined by events that occurred before the agents were born.[10]

Feltz and Cokely (2009) also adopt NCK's psychological, non-reductionistic description of determinism to probe intuitions about free will and moral responsibility, but add a fascinating twist: the authors investigate whether *personality differences* can affect intuitions on the compatibility question. Specifically, Feltz and Cokely predicted that subjects who were high in personal trait extroversion would be more likely to assign free will and moral responsibility to a murderer in the deterministic scenario (due to the subjects' increased sensitivity to the social features of a scenario). Rather remarkably, the results showed a significant correlation between extroversion and compatibilist judgments.[11] These results, if they can be replicated and expanded upon, have important implications for they suggest the possibility that intuitive differences related to free will and moral responsibility may not be resolvable by philosophical reflection and dialogue.

The final study I will discuss in this 'central dialectic' is presented in Roskies and Nichols (2009). The authors predicted that intuitions about free will and moral responsibility would be sensitive to whether deterministic scenarios are described as *actual*, in our world, or merely *possible* (true in some other possible world). Their study has two conditions. In the 'actual' condition, subjects receive the following deterministic scenario:

> Many eminent scientists have become convinced that every decision a person makes is completely caused by what happened before the decision – given the past, each decision has to happen the way that it does. These scientists think that a person's decision is always an inevitable result of their genetic makeup combined with environmental influences. So if a person decides to commit a crime, this can always be explained as a result of past influences. Any individual who had the same genetic makeup and the same environmental influences would have decided exactly the same thing. This is because a person's decision is always completely caused by what happened in the past (p. 3).

Subjects are then asked whether we can be free and morally responsible if these scientists are right about genes and environment completely causing our actions.

In the 'alternate condition', the description of determinism is the same, except that the universe the subjects are asked to consider is not our own. The subjects then respond to similar probes about free will and moral responsibility in this alternate universe. Consistent with the authors' hypothesis, the assignments of free will and moral responsibility were significantly higher in the actual condition than in the alternate condition. Roskies and Nichols use these results in part to undermine the claims of those who believe that widespread acceptance of determinism would have earth-shattering implications for

human affairs. Smilansky (2000), for example, argues that acceptance of determinism would deeply undermine our sense of moral worth and might even lead to an 'unprincipled nihilism'. On the optimistic end of the spectrum, Waller (1990), Pereboom (2001), Greene and Cohen (2004), and Sommers (2007) argue that it would have few negative effects and would have the positive effect of making us less retributive; we would understand that no one deserves blame and punishment and would evaluate our response to crimes and moral wrongdoing in a more practical fashion. Roskies and Nichols, however, believe their results to show that widespread acceptance of determinism would not have much of an effect at all, positive or negative, as the determinism would occur in our own world.

4. Are We Asking the Right Questions?

There is no doubt that experimental work on the compatibility question has made numerous contributions to our understanding of free will and moral responsibility. The studies have raised serious doubts about the folk's pretheoretic commitment to incompatibilism. They have illustrated the role of emotion in our moral responsibility judgments. They have revealed how factors such as personality differences, and the terms in which determinism is described (mechanistic or psychological, actual or possible) can influence our judgments on the compatibility question. These are important insights, and I do not wish to downplay their value. Nevertheless, there are some serious practical and philosophical worries about the approach adopted within this central dialectic. The aim of this section was to identify the source of these worries, in hopes that future experimentalists can find ways to address them.

I begin with the practical difficulties. The challenge of describing determinism to subjects unfamiliar with the concept is daunting to say the least. The description must: (1) make the determinism sufficiently salient, but (2) not trigger fatalistic interpretations, or (3) beg any questions about how to interpret words like 'can' and 'possibility' and terms like 'had to happen'. It is at least arguable that providing an unbiased non-technical description of determinism in half a page is an impossible task. (Indeed, when researchers give manipulation checks to test for comprehension of determinism, somewhere between 10% and 30% of the subjects have to be excluded.) Adding to the experimental noise are the different ways that subjects could interpret the concepts of free will and especially moral responsibility. The sense of moral responsibility at issue in the philosophical debate is a matter of some controversy, although most philosophers agree that it involves something like non-consquentialist desert. Getting this concept across in a scholarly article, never mind a survey, is extremely difficult. Nichols and Knobe, for example, ask whether agent can be 'fully morally responsible' in a deterministic world, a term that can mean a number of different things. Nichols and Roskies are a little more specific: they ask whether an agent 'should be morally blamed'. But that question has an available consequentialist interpretation under which even a hard determinist (or 'illusionist' like Smilansky) might answer in the affirmative. Other more precise terms like 'blameworthy' or 'deserves blame or praise' may be too technical for many subjects, although they probably can be described in ways that are accessible to a general audience. Of course, it is impossible to be *certain* that subjects comprehend the relevant concepts in acceptable, non-question-begging ways. (This is true of social psychology studies in general.) And a large part of the value of these studies involves asymmetries in responses that would be significant even if there were some lingering ambiguity about the concepts of determinism or moral responsibility. Nevertheless, as the philosophical debate is focused on the precise interpre-

tations of these concepts, we ought to devote as much effort as possible to clarifying them for the subjects – perhaps by explicitly ruling out common misinterpretations or by introducing 'briefing sessions' (see section 5).[12]

However, there is a deeper and more philosophical problem with directly probing intuitions about the compatibility question. The roots of this worry can be seen in NMNT's seminal article. In the section 'Why it matters that incompatibilism is intuitive', the authors observe quite rightly that arguments for incompatibilism always appeal to intuitions about key premises and cases, and that incompatibilist theories are ultimately be evaluated by how well they accord with our intuitions. Next, as noted above, they offer a series of quotations from incompatibilists who claim that the folk are natural pretheoretic incompatibilists. One of these is from Robert Kane who writes:

> In my experience, most ordinary persons start out as natural incompatibilists. They believe there is some kind of conflict between freedom and determinism; and the idea that freedom and responsibility might be compatible with determinism looks to them at first like a 'quagmire of evasion' (William James) or 'a wretched subterfuge' (Immanuel Kant). Ordinary persons have to be talked out of this natural incompatibilism by the clever arguments of philosophers (Kane 1999: 217; from Nahmias et al. 2006: 29).

The authors cite similar passages from Galen Strawson, Thomas Pink, and Laura Ekstrom. The motivation for the studies then seems clear. Kane et al. are making empirical claims about folk intuitions that can, and should, be tested. Subsequent articles in the experimental literature have followed NMNT in framing the issue in this manner, choosing from a menu of passages about the intuitiveness of incompatibilism and then providing data to support or undermine these claims. What's odd from a dialectical standpoint, however, is that the cited passages almost always come from *introductory* sections of incompatibilist articles or books. The role of these passages is largely rhetorical, to get readers into an incompatibilist mood. There is no incompatibilist *argument* which features a premise like 'free will and moral responsibility are incompatible with determinism' (for good reason as this would be an obvious case of question-begging). Nor are there arguments of the form: (1) we intuitively find free will and moral responsibility to be incompatible with determinism, and therefore (2) free will and moral responsibility are incompatible with determinism. So, while NMNT (and other experimentalists who have adopted their approach) are correct that incompatibilist arguments appeal to intuitions, the appeal does not directly concern the question that is probed in their studies.

What, then, are the implications of NMNT's results for incompatibilism? The authors write:

> Either determinism obviously precludes free will or those who maintain that it does should offer an explanation as to why it does. The *philosophical* conception of determinism – i.e., that the laws of nature and state of the universe at one time entail the state of the universe at later times – has no obvious conceptual or logical bearing on human freedom and responsibility. So, by claiming that determinism *necessarily* precludes the existence of free will, incompatibilists thereby assume the argumentative burden (Nahmias et al. 2006, 30).

NMNT believe that their results shift the burden of proof to the incompatibilist. It seems, however, that incompatibilists can accept that they have this 'argumentative burden' but claim that they have discharged it with, well, arguments. After all, van Inwagen's 'consequence argument', Strawson's 'basic argument', and Pereboom's 'four case argument', to name just a few, are designed precisely to lead the reader to the conclusion that determinism precludes free will and moral responsibility. Again, these arguments do

appeal to intuitions, but to intuitions about cases and narrower principles (such as the PAP or the 'transfer of non-responsibility principle') rather than on the compatibility question itself. It would seem that in order to truly test the plausibility of the incompatibilist position, we need to examine the intuitions support the premises and principles of their arguments and not their introductory rhetoric.

NMNT do anticipate an objection of this kind, but their response is problematic. Noting that they have not tested something like the consequence argument directly, they reply that their results provide *indirect* evidence against its intuitive plausibility because '[o]ur scenarios present conditions in the past that, along with the laws of nature, are sufficient conditions for the agent's action'. (45) However, this reply assumes that philosophically unsophisticated subjects should have somehow internalized or implicitly grasped the consequence argument before ever hearing it! Van Inwagen has received credit for reviving incompatibilism in the past century in large part because he found a non-obvious way of showing the threat of determinism to free will. Indeed, the whole point of developing an incompatibilist argument (rather than just asserting the conclusion) is to draw out implications of determinism that might otherwise go unnoticed. So it is not clear that NMNT's results give much in the way of indirect evidence against incompatibilist arguments either.[13]

I have focused my remarks on NMNT because they developed the first of these studies, and because their article features the most comprehensive defense of the methodology. But the objection applies to Nichols and Knobe (2007), Nahmias et al. (2007), and all of the subsequent studies that employ this approach. One might wonder, then, why we should care at all about these studies as they do not probe intuitions about the premises of incompatibilist or compatibilist arguments?[14] The answer is that even if my critique in this section is accurate, the work described will have offered crucial insights into the psychology underlying our judgments about free will and moral responsibility. It is likely that the factors uncovered by earlier studies such as affect, personality traits, and description type have important effects on intuitions about the principles and cases at the core of the philosophical debate. I offer these criticisms in hopes that experimentalists can build on the insights of their predecessors by developing new experiments to probe these intuitions.

5. Other Experimental Work on Free Will

Although the compatibility question has drawn the most attention from experimental philosophers, there is a good deal of work that addresses other aspects of the free will debate as well. Several recent studies, for example, have attempted to shed light on the consequences of *believing* that there is no free will. As noted above, many philosophers have argued that accepting the hard determinist thesis would have dire implications for our sense of self-worth, and might even incline us to behave unethically. Smilansky (2000) has gone so far as to suggest that we should keep the *illusion* of libertarian free will in place, so as not to bring out these unfortunate consequences. Until recently, however, there was no empirical evidence to support these pessimistic claims. This changed with the appearance of Vohs and Schooler (2008) a study that was pounced upon by the popular press. (See here for one example.) Subjects in the study were divided into two groups. In the first condition, subjects read a passage from Francis Crick's *The Astonishing Hypothesis* which stated that scientists had denounced the notion of free will. In the second condition, subjects read another excerpt from Crick's book about consciousness, but one that did not refer to free will. The authors then gave the subjects cognitive tasks

which featured opportunities to cheat, including one where they could pocket some extra money for doing so. As predicted, the subjects who had read the denunciation of free will were more likely to cheat on their tasks. Vohs and Schooler conclude their paper by asking: 'Does the belief that forces outside the self determine behavior drain the motivation to resist the temptation to cheat, inducing a ''why bother'' mentality? … Or perhaps denying free will simply provides the ultimate excuse to behave as one likes'.[15]

Experiments like these provide an important challenge to the 'happy hard determinists' (Smilansky's term) who argue that denying free will would have beneficial implications on the whole. However, do the results show that Smilansky's worst fears are realized? I do not believe so, for it is not clear that our behavior just after hearing that a cherished belief is false has any bearing on how we would act after further reflection.[16] The authors cite no studies that document such a correlation. To put the point somewhat contentiously, Vohs and Schooler may merely have shown that people should not become hard determinists 15 min before they submit their tax return. The philosophical debate focuses on the long-term implications of a widespread denial of free will.

Nadelhoffer and Feltz (2007) offer additional reasons to question the pessimism of Smilansky. They note, correctly, that testing Smilanksy's claims is extremely difficult as people are almost certainly unreliable predictors about how they themselves and others would react if they lost their belief in free will.[17] Ideally, we could observe the behavior of people who have denied free will and moral responsibility for long periods. Unfortunately, those people are few and far between. However, as Nadelhoffer and Feltz note, Smilansky *himself* seems to lead an honorable and gratifying life in spite of being disillusioned about free will. The same goes for skeptics such as Pereboom, Galen Strawson, Bruce Waller, and by all accounts, Denis Diderot and Baruch Spinoza. Why should we suspect that the rest of humanity would suffer long-term harmful consequences?[18]

Another subject for experimental inquiry has been the role of alternate possibilities and identification in moral responsibility judgments. Incompatibilists have traditionally appealed to two principles in their arguments, one of which is known as the PAP. According to PAP, agents cannot be morally responsible for an act if they cannot do other than perform that act. If PAP is true, and determinism entails that we can only perform actions that we in fact perform, then the incompatibilist conclusion follows.

Frankfurt (1969) offers a famous counterexample to PAP, which may be summarized as follows. Imagine that an evil neurosurgeon has placed a chip inside the brain of an assassin who is considering whether to shoot a prime minister. The chip allows the neuroscientist to witness the deliberations of the assassin and, if necessary, to alter the assassin's behavior. If the assassin decides to go through with the plan, the neuroscientist will do nothing as he wants the prime minister dead. However, if the assassin has second thoughts and decides not to fire, the neuroscientist will press a button which compels the assassin to follow through with his assignment. The scenario appears to be a counterexample to PAP because: (1) the assassin cannot do otherwise than shoot the prime minister, but (2) *if* the assassin decides on his own to go through with the killing, he will still be morally responsible for doing so. Building on this counterexample, Frankfurt argues that what really matters for responsibility is not whether the agent could have done otherwise, but rather whether the agent *identifies* with the act, endorses it with a higher order desire or volition. (Frankfurt 1971, 1988).

Two recent studies appear to offer support for Frankfurt's theory. In the first, Woolfolk et al. (2006) devise a scenario in which an agent, Bill, is captured by hijackers on an airplane and ordered to kill Frank, whom Bill knows to be his wife's lover. The hijackers give Bill a 'compliance drug' which makes it impossible for him to disobey. In the 'low

identification condition', subjects are told that Bill, although not thrilled about the cuckolding, is appalled by this act and only performs it because he is forced to by the compliance drug. Subjects in the 'high identification condition' are told that Bill is grateful for the opportunity to kill his wife's lover with impunity and feels absolutely no reluctance about performing it. The results show the effect of identification: subjects in the high identification condition assigned higher ratings of responsibility to Bill than subjects in the low identification condition. A second study, by J. Miller and A. Feltz (submitted), offers further support for Frankfurt. The authors gave subjects Frankfurt-style cases in which an agent is compelled either to perform an immoral action or to refrain from preventing an immoral act from occurring. In both conditions, subjects were more likely to attribute responsibility to agents who had decided on their own to perform (or refrain from preventing) the immoral act.

Both of these studies succeed in: (a) showing the importance of identification to responsibility, and (b) raising doubts about the PAP. They do not, however, go very far towards vindicating compatibilism about responsibility as they leave the second and much stronger incompatibilist principle, the 'transfer of non-responsibility (TNR) principle', untouched. Roughly, the TNR principles asserts that if an agent is not even partially morally responsible for any of the factors that were sufficient to cause an act (where determining factors may include acts of omission), then the agent cannot be morally responsible for the act itself.[19] The TNR principle, or relatives of it (like Van Inwagen's 'Rule B'), is at the heart of almost all recent incompatibilist arguments about responsibility, including Van Inwagen (1983), Strawson (1986), Kane (1996), and Pereboom (2001). Indeed, Kane allows for agents to be blameworthy for fully determined actions as long as they are ultimately responsible for their characters. One plausible interpretation of Frankfurt-style scenarios is that agents in high identification conditions are held responsible for their immoral characters. However, perhaps, like Kane, subjects believe that in order to be responsible for one's character, one must have made an indeterministic 'self-forming' choice earlier in life that led them to have this immoral character. They would then reject PAP while remaining incompatibilist. These considerations point to the importance of developing studies to probe intuitions about the TNR principle.

For space reasons, I can do no more than direct the reader to several additional studies (but the reader should note that this is a burgeoning field with new work appearing all the time). Nichols (2004), arguably the first real piece of experimental philosophy on free will, presents evidence suggesting that children take themselves to be agent causes, and canvasses some possible explanations for how children acquire this belief.[20] Sarkissian et al. (2009) examines intuitions about the compatibility question across cultures, and finds, rather surprisingly, that there is not so much variation, at least when questions are probed abstractly.[21] The study will hopefully mark a new trend within the field as cross-cultural research is, in my view, an especially fertile ground for further research. Finally, Morris (in preparation) is developing studies to test intuitions about whether agents in a deterministic universe have the ability to do otherwise. If successful, these studies could make progress in resolving the dispute between Nahmias and Nichols and Knobe about how subjects are interpreting descriptions of determinism in which actions and decisions 'had to happen'.

6. Avenues for Further Investigation

In conclusion, I would like to offer some brief suggestions to address some of the worries raised in this article, and also to point to some unexplored territory in the free will terrain.

6.1. DEVELOP STUDIES THAT TARGET INCOMPATIBILIST PRINCIPLES AND CASES

I have argued at some length that future experimental work should test the intuitiveness of incompatibilist principles rather than the incompatibilist conclusion. Experiments examining intuitions about the cases in manipulation arguments (such as the 'Zygote argument', Mele 2006; or the 'Four Case Argument', Pereboom 2001) might be valuable in this respect. More generally, experimentalists might do well to focus on the specific conditions purported to be necessary and/or sufficient for free and responsible action and move away from the compatibility question entirely.

6.2. HOLD BRIEFINGS SESSIONS THAT HELP TO CLARIFY THE CONCEPTS LIKE DETERMINISM, MORAL RESPONSIBILITY, AND FREE ACTION[22]

As noted in section 4, the concepts of moral responsibility, determinism and free will are open to many interpretations, including some that are not subject to much philosophical controversy. There are legitimate concerns about whether subjects conflate determinism and fatalism, moral, and legal responsibility, and whether they give consequentialist responses to questions about blame and praise rather than focusing on desert. One beguilingly simple way of addressing these problems is to hold 30- to 60-min sessions in which the subjects are given precise (non-question-begging) accounts of the concepts at issue, and offered the opportunity to ask questions and clear up confusions. This would probably require that the subjects be paid a small amount to participate. However, if the sessions generated greater confidence in the results, this might well be worth the cost. Of course, these 'training' sessions would raise different worries about biasing subjects; so, it is crucial to find neutral ways of operationalizing the trainings sessions that would satisfy proponents of both sides of the debate. However, if done correctly, this method might help to answer critics who focus on the 'prereflective' nature of the judgments experimental philosophers probe.[23]

6.3. USE REAL-WORLD SCENARIOS AND VARY THE PSYCHOLOGICAL DISTANCE OF THE OFFENSES

One considerable virtue of Roskies and Nichols (2009) article is their demonstration that our intuitions vary according to how realistic and 'close to home' the scenarios are presented. As they note, their results may 'call into question traditional philosophical reliance on possible worlds analyses for understanding our concepts' (16). As we are interested in free will and moral responsibility in the *actual* world, scenarios about these concepts should perhaps be restricted to events that occur within it. These considerations also speak against using exceedingly unrealistic descriptions of events in our own universe. (We might be more judicious about the use of nefarious neurosurgeons and omniscient supercomputers, for example.)

In addition, one might build on Roskies and Nichols' insights by probing intuitions about scenarios that grow increasingly closer to the subject's life and experiences. Intuitions about freedom and responsibility may be altered when the acts occur in the subject's home state, or home town, rather than in another country.[24] Intuitions may also be sensitive to whether the offender, or victim, is related to the subject. If it turns out that intuitions vary according to psychological distance and/or personal relationship, this might support 'variantist' theories of moral responsibility (Doris et al. 2007; Knobe and Doris 2009) in which different criteria for responsibility apply within different contexts.

6.4. DEVELOP BEHAVIORAL EXPERIMENTS

As I suggest above (note 1), experiments that are not designed to probe intuitions may still provide important insights into philosophical questions. Behavioral experiments are a prime example. It is one thing to ask subjects whether a man was compelled to behave immorally is morally responsible, another to offer them the opportunity to punish the offender at a cost. Experiments employing public goods, games, ultimatum games, and prisoner dilemmas might offer a fascinating glimpse into how our behavior is affected by factors such as determinism, identification, and intention. It seems quite possible that our judgments might not always translate into neatly corresponding behavioral responses. Uncovering the relationship between judgment and behavior would markedly contribute to our understanding of what we really think about freedom and responsibility.

Of course, there are numerous other exciting ways of improving existing paradigms and creating new ones – some have no doubt already been developed. The good will that exists among experimentalists and more traditional analytic philosophers on this topic should lead to a productive future, with a variety of methods combining to provide a deeper understanding of free will and moral responsibility.

Short Biography

Tamler Sommers is Assistant Professor of Philosophy at the University of Houston and holds a joint appointment in the Honors College. He has published articles in *The Philosophical Quarterly*, *Biology and Philosophy*, *Philosophy and Phenomenological Research*, and *Psyche*. He is currently writing a book on perspectives about moral responsibility across cultures entitled *Relative Justice* (under contract, Princeton University Press). His book of interviews, *A Very Bad Wizard: Morality Behind the Curtain*, will be published in October, 2009 by McSweeney's Press.

Acknowledgements

I am grateful to Ron Mallon, Eddy Nahmias, and an anonymous referee at Philosophy Compass for valuable comments on earlier drafts of this paper. Portions of this paper were written at the NEH Summer Institute on Experimental Philosophy. I thank the directors Ron Mallon and Shaun Nichols for giving me the opportunity to attend the program.

Notes

* Correspondence: Tamler Sommers, 513 Agnes Arnold Hall, Houston, TX 77004, USA. Email: tamlers@gmail.com

[1] Indeed, some philosophers think that the term 'experimental philosophy' should only apply to work that attempts to probe our intuitions and judgments. In my view, this definition is overly restrictive. For one thing, it would exclude behavioral experiments (such as Vohs and Schooler 2008; described below) that are designed to address concerns at the center of philosophical debates. See also Nadelhoffer and Nahmias (2007) for an excellent overview of the diverse aims of experimental philosophy.

[2] Knobe and Nichols (2008, p. 3).

[3] See section 5 for a more detailed description of these principles.

[4] Consider Van Inwagen's defense of beta, the 'transfer of powerlessness' principle which is central to his 'consequence argument' for incompatibilism about free will:

I must confess that my belief in the validity of beta has only two sources, one incommunicable and the other inconclusive. The former source is what philosophers are pleased to call 'intuition'… The latter source is the fact

that I can think of no instances of Beta that have, or could possibly have, true premises and a false conclusion (Van Inwagen 1983, pp. 97–99).

[5] See Sommers (2009) for further discussion of how incompatibilists must appeal to intuition to support their arguments.

[6] Frankfurt's (1969) famous case of the assassin, for instance, can only be a successful counterexample to the PAP if we find the assassin intuitively blameworthy when there is no interference. Fischer and Ravizza employ the same strategy in their arguments against the 'transfer of non-responsibility' principle (see Ravizza 1994; Fischer and Ravizza (1998); and the 'erosion' and erosion* cases).

[7] Although the details of Frankfurt and Watson's accounts differ, they share the basic view that we can be morally responsible for actions that stem are reflectively endorsed by a deeper self than a mere first-order desire.

[8] For more detailed defenses of the relevance of experimental philosophy on free will and other topics, see Nahmias et al. (2006), Alexander and Weinberg (2006), Nadelhoffer and Nahmias (2007), Knobe and Nichols (2008, Chapter One). For a thoughtful critique, see Kauppinen (2007).

[9] It is worth noting that NMNT do a fair amount of 'armchair' philosophizing to reach this conclusion. I will discuss their argument in section 3.

[10] Nahmias and Murray (2009) have recently conducted a study that more straightforwardly examines the possibility of bypassing interpretations. After giving subjects Nichols and Knobe's description of determinism, they ask subjects whether the agent's action could have been different if the desires, intentions, and thoughts of the agent had been different. The results suggest that subjects may be conflating fatalism with determinism, assuming that agents would perform their actions whether or not they had the intention/desire to do so.

[11] It is worth noting that these results may offer support for Strawson (1962) and Watson's (1987) claims that judgments about free will and moral responsibility are essentially tied to social attitudes and practices. Watson notes that Einstein, a hard determinist, describes himself as a 'lone traveler' in the world.

[12] Thanks to an anonymous referee for highlighting the importance of this worry. See also Feltz et al. (2009) for a penetrating critique of the studies discussed in section 2.

[13] NMNT offer another perhaps more plausible reason for incompatibilists to worry about their results. They write that even if the folk find the premises of the consequence argument to be intuitive, this would merely show that they have conflicting intuitions about the threat of determinism to freedom and responsibility. Incompatibilists would then have to show that an intuition about a transfer principle, say, is more basic than an intuition about the incompatibility question itself.

[14] Thanks to an anonymous referee for encouraging me to respond to this worry.

[15] Baumeister et al. (2009) have expanded upon these results with a study which suggests that inducing disbelief in free will leads to an increase in aggression and a reduction in willingness to help (p. 267). They also tested for 'chronic disbelief' in free will and observed a correlation between subjects who were high on this scale with lack of helpfulness. In line with Nahmias et al. (2007), however, many of the questions in this probe test conflate fatalistic beliefs with beliefs about free will.

[16] Having one's entire worldview shaken up can certainly have some short-term negative effects on behavior. Fifteen minutes after Grady Little inexplicably allowed Pedro Martinez give up the lead to the Yankees in game 7 of the 2003 ALCS, I might have committed acts that are far more immoral than cheating on a psychology quiz.

[17] Nadelhoffer and Feltz also run a study which asks one set of subjects how *they* would behave if they were to learn that we lack free will, and another how *others* would behave upon acquiring this knowledge. Whereas only 35% of subjects believed that they would behave immorally, 70% believed that others would do so. According to the authors, these results show that we are not unreliable predictors of how we would behave in light of new knowledge.

[18] Indeed, inspired by Feltz and Cokely, I suspect that personality differences would have an enormous effect on how one responded to a belief in hard determinism. However, it is not clear to me how we to develop an effective empirical test for this prediction.

[19] Fischer and Ravizza (1998) provide the following more careful formulation of this principle: '(1) p obtains and no one is even partly morally responsible for p; and (2) if p obtains, then q obtains, and no one is even partly morally responsible for the fact that if p obtains, then q obtains; then (3) q obtains, and no one is even partly morally responsible for q' (p. 152). This 'non-responsibility' for the original factors that produced the act is transferred to the act itself.

[20] See Turner and Nahmias (2006) for a response.

[21] Subjects were not asked about concrete cases in this study.

[22] I owe this suggestion to Jonathan Weinberg.

[23] Thanks to Ron Mallon for encouraging me to emphasize this point.

[24] Chris Wiegel is currently devising experiments to test the effects of psychological distance on judgments regarding the compatibility question.

Works Cited

Alexander, J. and Weinberg, J. 'Analytic Epistemology and Experimental Philosophy.' *Philosophy Compass* 2.1 (2006): 56–80.

Baumeister, R.F., Masicampo, E.J. and DeWall, C.N. 'Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness.' *Personality and Social Psychology Bulletin* 35 (2009): 260–2.
Doris, J, Knobe, J. and Woolfolk, R. 'Variantism about Responsibility.' *Philosophical Perspectives* 21 (2007): 183–214.
Feltz, A. and Cokely, E.T. 'Do Judgments about Freedom and Responsibility Depend on Who You Are?: Personality Differences in Intuitions about Compatibilism and Incompatibilism.' *Consciousness and Cognition* 18 (2009): 342–50.
——, Cokely, E.T. and Nadelhoffer, T. 'Natural Compatibilism v. Natural Incompatibilism.' *Mind and Language* 24 (2009): 1–23.
Fischer, J. and Ravizza, M. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, MA: Cambridge University Press, 1998.
Frankfurt, H. 'Alternate Possibilities and Moral Responsibility.' *Journal of Philosophy* 66 (1969): 829–39.
——. 'Freedom of the Will and the Concept of a Person.' *The Importance of What We Care About*. Cambridge, MA: Cambridge University Press, 1971, 1988. 11–25.
Greene, J. and Cohen, J. 'For the Law, Neuroscience Changes Nothing and Everything.' *Philosophical Transactions of the Royal Society of London B* ccclix (2004): 1775–85.
Kane, R. *The Significance of Free Will*. Oxford: Oxford University Press, 1996.
——. 'Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism.' *Journal of Philosophy* 96/5 (1999): 217–40.
Kauppinen, A. 'The Rise and Fall of Experimental Philosophy.' *Philosophical Explorations* 10.2 (2007): 95–118.
Knobe, J. 'Experimental Philosophy and Philosophical Significance.' *Philosophical Explorations*, 10 (2007): 119–22.
—— and Doris, J. 'Strawsonian Variations.' *The Handbook of Moral Psychology*. Eds. J. Doris et al. Oxford: Oxford University Press, 2009, in press.
—— and Nichols, S. 'Experimental Philosophy Manifesto.' *Experimental Philosophy*. Eds. J. Knobe, S. Nichols. Oxford: Oxford University Press, 2008. 3–14.
Mele, A. *Free Will and Luck*. Oxford: Oxford University Press, 2006.
Nadelhoffer, T and Feltz, A. 'Folk Intuitions, Slippery Slopes, and Necessary Fictions: An Essay on Saul Smilansky's Illusionism.' *Midwest Studies in Philosophy* 31 (2007): 202–13.
—— and Nahmias, E. 'The Past and Future of Experimental Philosophy.' *Philosophical Explorations* 10.2 (2007): 123–49.
Nahmias, E. and Murray, D. 'Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions.' *New Waves in Philosophy of Action*. Eds. J. Aguilar, A. Buckareff, K. Frankish. New York: Palgrave-Macmillan, 2009, Forthcoming.
——, Morris, S., Nadelhoffer, T. and Turner, J. 'Surveying freedom: Folk intuitions about free will and moral responsibility.' *Philosophical Psychology* 18 (2005): 561–84.
——, ——, —— and ——. 'Is Incompatibilism Intuitive?' *Philosophy and Phenomenological Research* 73 (2006): 28–53.
——, Coates, D. and Kvaran, T. 'Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions.' *Midwest Studies in Philosophy* 31 (2007): 214–42.
Nichols, S. 'Folk Psychology of Free Will.' *Mind and Language* 19 (2004): 473–502.
—— and Knobe, J. 'Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions.' *Nous* 43 (2007): 663–85.
Pereboom, D. *Living Without Free Will*. Cambridge, MA: Cambridge University Press, 2001.
Roskies, A.L. and Nichols, S. 'Bringing Moral Responsibility Down to Earth.' *Journal of Philosophy* 2009, forthcoming.
Sarkissian, H., Chaterjee, A., De Brigard, F., Jelly, C., Knobe, J., Nichols, S. and Sirker, S. 'Is Belief in Free Will a Cultural Universal?' *Mind and Language*, Forthcoming.
Smilansky, S. *Free Will and Illusion*. Oxford: Oxford University Press, 2000.
Sommers, T. 'The Objective Attitude.' *The Philosophical Quarterly* 57.28 (2007): 321–42.
——. 'More Work for Hard Incompatibilism.' *Philosophy and Phenomenological Research* LXXIX.3 (2009): 511–21.
Strawson, P.F. 'Freedom and Resentment.' *Proceedings of the British Academy* 48 (1962): 1–25.
Strawson, G. *Freedom and Belief*. Oxford: Clarendon Press, 1986.
Turner, J. and Nahmias, E. 'Are the Folk Agent Causationists?' *Mind and Language* 21.5 (2006): 597–609.
Van Inwagen, P. *An Essay on Free Will*. Oxford: Clarendon Press, 1983.
Vohs, K.D. and Schooler, J.W. 'The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating.' *Psychological Science* 19 (2008): 49–54.
Waller, B. *Freedom without Responsibility*. Philadelphia: Temple University Press, 2009.
Watson, G. 'Responsibility and the Limits of Evil.' *Responsibility, Character, and the Emotions*. Ed. F. Schoeman. Cambridge, MA: Cambridge University Press, 1987. 256–86.
Wolf, S. 'Sanity and the Metaphysics of Responsibility.' *Responsibility, Character and the Emotions: New Essays in Moral Psychology*. Ed. F. Schoeman. Cambridge, MA: Cambridge University Press, 1987. 46–62.
Woolfolk, R., Doris, J. and Darley, J. 'Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility.' *Cognition* 100 (2006): 283–301.