Experimental philosophy and moral responsibility

GUNNAR BJÖRNSSON, STOCKHOLM UNIVERSITY

Experimental philosophy, as I understand it here, is the attempt to help answer ontological, semantic, or epistemological questions about some matter of philosophical concern by quantitatively measuring judgments about that subject matter, typically among non-philosophers.

The last couple of decades has seen a flurry of such attempts in various subdisciplines, including epistemology, philosophy of mind, philosophy of language, and action theory. Philosophers have of course always relied on judgments about their subject matters, but have less often relied on non-philosopher's judgments, and rarely in ways involving quantitative assessments of the distribution of such judgments. Philosophers have also from time to time relied on quantitative empirical data, perhaps most prominently in philosophy of mind and applied ethics, but rarely on data on folk judgments about the subject matter of concern. Experimental philosophy contrasts with both these more traditional philosophical practices.

The ambitions of experimental philosophers vary. Some modestly hope to correct misapprehensions about how certain expressions are ordinarily used. Others think that empirical methods might radically reshape philosophical inquiry. The revolutionary potential has prompted interesting broad questions: Might experimental philosophy show that attempts to solve philosophical problems "from the armchair" are fundamentally mistaken, or are data on judgments largely irrelevant for central philosophical problems?[1] Under what conceptions of philosophical inquiry might empirical methods be more or less relevant? Are some philosophical questions particularly amenable to these empirical methods, whereas others are not? Should experimental philosophy both become the sustained concern of a substantial group of specialists and a source of relatively accessible tools used throughout philosophy, much as philosophical logic?[2]

Some of what is said in this chapter is relevant to these broader issues. Our question is narrower, however, in keeping with the focus of this volume: How, if at all, might experimental philosophy help us answer central questions about the nature of moral responsibility, such as the question of whether moral responsibility is compatible with determinism?[3] The primary focus will be on the idea that folk judgments in line with a particular position might provide

[1] For discussion, see e.g. Kauppinen 2007; Knobe 2007; Alexander et al. 2010 and contributions to Sytsma and Buckwalter 2016.

[2] For the last suggestion, see Weinberg 2015. For some recent overviews of the problems and prospects of experimental philosophy, see Mallon 2016; Knobe and Nichols 2017; Ludwig 2017. For data on the reproducibility of results in experimental philosophy, see Cova et al. 2018; for discussion of the use of online subjects, see e.g. Paolacci and Chandler 2014.

[3] For some earlier overviews of the relevance of experimental philosophy to the philosophy of moral responsibility specifically, see Vargas 2006; Nelkin 2007; Sommers 2010; Björnsson and Pereboom 2016; Chan et al. 2016.

support for that position. But we will also consider studies suggesting that certain classes of judgments are particularly unreliable, and ask whether experimental results are relevant for our understanding of philosophers' judgments about the subject matter.

Throughout these discussions, I will assume a broad familiarity with philosophical research on moral responsibility. Understanding the philosophical relevance of experimental results involves the same difficulties and requires the same familiarity with distinctions and possible theoretical options as does understanding the relevance of arguments that rely, in familiar ways, on philosophers' verdicts about cases.

1. Wisdom of the crowd?

Why should philosophers care about what non-specialists think about a complex philosophical question? A straightforward answer would be this: if one position on this issue is supported by significantly more people, this provides some evidence for that position. In line with this, some of the early experimental work on whether non-philosophers are compatibilists or incompatibilists was indeed a response to claims by some philosophers that their position had common sense on its side, claims taken (correctly or not) to imply that the burden of proof would be on the other side (Nahmias et al. 2005: 563–4). Some will balk at this answer. Given the difficulty of the questions about incompatibilism and given how easily people get confused when thinking about it, one might find it obvious that folk judgments carry no evidential weight whatsoever. But there is reason to at least take the suggestion seriously and be clear about why we reject it.

Generally speaking, and for most of the questions we consider in our lives, our ordinary ways of answering these questions are somewhat reliable, or better than chance at getting things right. But if people in general are slightly more likely than not to get things right, and if their chances of doing so are independent, then if a significant majority of a large number of people agree on a given issue, the majority would most likely be correct. Or at least they would be in cases where there is one correct answer, no answer is antecedently more likely than the other, and judges would be equally likely of getting things right independently of which answer is in fact correct.[4] Under those conditions, if the chance for each to arrive at the correct answer is a modest .55, then if 61 out of 100 such judges come down on one side, there is nevertheless over .99 chance that they are correct. If the chance for each is a mere .505, then the same is true if it is supported by 605 out of 1,000 judges.[5] The sheer difference in number of supporting judgments make up for the modest or meagre quality of each judgment. How is this possible?

[4] The reasoning that is merely roughly indicated here is made more precise in discussions of the Condorcet Jury Theorem. For a discussion of the relevance of CJT to experimental philosophy and a more formal treatment of both reliability and independence, see Talbot 2014; Ludwig 2017. Cf. Goodin and Spiekermann 2018.

[5] These numbers, and corresponding numbers provided later on are calculated based on Theorem 9 in List 2004: 528. Cf. Talbot 2014. I move freely between talk of probabilities, likelihoods, and chances. What I have in mind in each case is something like *fitting credences given accessible information*; I take the discussion to be robust relative to various ways of spelling out that notion.

It helps to think of marginally reliable judges as being mostly affected by factors that do not track the truth of the matter—by *noise*—and only to some small degree by factors that do. While noise is random and cancels out over a large number of independent judgments, the rare instances where judgments are guided by the truth always pull in the same direction.

The strength of this sort of appeal to the wisdom of the crowd depends on the extent to which:

(1) *No answer is antecedently more likely* than others in a way that lets us significantly discount the reliability of judges arriving at some other answer.

(2) *People's chances of getting things right are independent*. Suppose that each is more likely than not to get things right but is so by leaning on one and the same piece of evidence, which is slightly more likely than not to reveal the truth. Then even if noise cancels out, uniformity in judgment will primarily reflect what that piece of evidence suggests. However large the group and uniform their agreement, the likelihood that the group gets things right will not exceed the likelihood that the evidence in question is veridical.

(3) *People are reliable when judging this particular issue*, being somewhat better than chance in getting things right. (We obviously cannot assume that people are highly reliable when it comes to difficult philosophical questions. But given that even very low reliability can ground a wisdom of the crowd argument, we cannot reject this reliability assumption without further consideration.)

(4) *The question has one correct answer*, independently of who is answering it. If judgments are tracking different facts, raw numbers cannot be expected to accumulate truth-tracking while noise cancels out.

(5) *A large enough majority of judgments* comes down on one side of this issue compared to the other. If the majority is very slim, there is little accumulation of evidential weight.

Can such an argument be had for central questions about responsibility? In what follows, I will discuss each condition, in reverse order. To constrain the discussion, I will continue to focus primarily on the question of whether incompatibilism is correct and restrict myself to a limited range of studies telling us whether and under what circumstances people tend to make compatibilist and incompatibilist judgments, respectively. Though this will mean ignoring other interesting studies, I take large parts of what is said here to generalize.[6] For the same reason, I will also assume (unless otherwise stated) that compatibilism and incompatibilism are antecedently equally likely.

[6] Among studies that I set to the side are those concerned with the extent to which people acknowledge moral luck (e.g. Cushman 2008; Kneer and Machery 2019), how people understand the relation between consciousness and responsibility (e.g. Shepherd 2015), and whether people take there to be historical conditions on responsibility (Taylor and Maranges Forthcoming).

2. How many accept or reject compatibilism?

Several studies have tried to determine whether non-philosophers take determinism to be incompatible with moral responsibility; whether they are "natural incompatibilists" (e.g. Nichols 2004; Nahmias et al. 2005; Nahmias et al. 2007; Nichols and Knobe 2007; Roskies and Nichols 2008; Sarkissian et al. 2010). In these studies, subjects are typically presented with either an indeterministic or a deterministic scenario and asked about whether agents can be morally responsible for their decisions or actions in that scenario. If judges are willing to attribute responsibility in the former but not the latter scenario, this would suggest that they are incompatibilists.

What has been the conclusion of these studies? Unfortunately, results have varied, depending on how the (in)deterministic aspect of the scenarios were described and on how questions about responsibility were asked. For example, looking at two of the most cited studies, the proportion of subjects that are willing to attribute responsibility to agents in deterministic scenarios vary between 14% (Nichols and Knobe 2007: 670) and 89% (Nahmias et al. 2007: 227). A number of reasons for this variation have been identified:

> *Abstract vs. Concrete*: When subjects are asked to attribute responsibility to a specified individual for a concrete bad action in a deterministic universe ("Is it possible that Bill is fully morally responsible for killing his family?"), they attribute more responsibility than when asked, in the abstract, whether someone could be fully morally responsible in such a universe ("Is it possible for a person to be fully morally responsible for their actions?"). In one study, 72% of subjects who were asked the concrete question gave a positive answer whereas only 14% of subjects who were asked the abstract question (Nichols and Knobe 2007); in another, the proportion of positive answers fell from 79% to 52% (Nahmias et al. 2007).
>
> *Actual vs. Possible*: When subjects are asked to consider whether agents would be fully morally responsible if the actual world is deterministic, they are also more inclined to give a positive answer than when considering agents located in some merely imagined deterministic universe. In one study, 89% gave positive answers to the former question, but only 72% to the latter (Nahmias et al. 2007; cf. Roskies and Nichols 2008).
>
> *Neurology vs. Psychology*: Subjects seems considerably more inclined to attribute responsibility when the deterministic scenario describes the deterministic processes in psychological terms (minds, thoughts, desires, plans) than when they are described in neurological terms (brains, chemical reactions, neural processes): 87% versus 41% in one study (Nahmias et al. 2007; but see De Brigard et al. 2009; Chan et al. 2016).
>
> *Framing of determinism*: Subjects are significantly more inclined to attribute responsibility when the idea of determinism is conveyed as the idea that the same events *always follow causally* from the same initial conditions than when put in strongly modal terms ("given the past, each decision *has to happen* in the way that it does"): 52% versus 27% in one study (Murray and Nahmias 2014: 447).

Some of these variations suggest that certain ways of asking or framing the question change what answers seem reasonable without therefore changing the question (see e.g. Nelkin 2007; Björnsson and Persson 2013, apart from discussion in papers presenting the original studies). We will return to matter of whether these variations affect the reliability of the answers given.

Other variations might suggest that subjects have been asked questions that have different answers. Most obviously, it might be that different ways of conveying the deterministic aspect of the scenarios lead subjects to consider different versions of determinism, only some of which are incompatible with responsibility. More subtly, it might be that the nature of responsibility depends on how things are in the actual world, and that this might explain the *actual vs. possible* contrast (Roskies and Nichols 2008: 382–5). For example, it could be that responsibility should be understood as requiring a kind of freedom ruled out by determinism if, as most people seem to believe (Nichols and Knobe 2007; Sarkissian et al. 2010; Deery et al. 2013), human decisions and actions are actually free in that way. But if the world is deterministic, as subjects are asked to assume in questions about the actual world, then perhaps responsibility should be understood in a way that still takes us to be responsible: the responsibility assumption would trump incompatibilist commitments (cf. van Inwagen 1983: 219ff).

Another equivocation worry stems from the well-known ambiguity of the term "responsibility". Even if we restrict ourselves to retrospective moral responsibility, some have argued that it comes in several importantly different varieties, sometimes labeled "accountability", "attributability", and "answerability" (see e.g. Watson 1996; Shoemaker 2015; for criticism, see Smith 2015; Jeppsson 2016), where accountability for an action would be what is required for an agent to deserve retributive emotional and other responses. One possible explanation of diverging responsibility judgments about deterministic cases might then be that whereas accountability is generally taken to be incompatible with determinism, some subjects have other, less demanding, notion of responsibility in mind, one that merely requires that it makes sense to ask the agent why they did what they did, demand that they abstain from such actions, or assess their character based on their action (Björnsson and Pereboom 2016: 152). There is some reason to think that this cannot explain more than part of the variation in judgment, however. Several studies of folk judgments ask not only about responsibility but also whether agents deserve blame. Since answers to these two questions tend to be fairly strongly correlated, there is reason to think that subjects are concerned with responsibility understood exactly as a precondition for deserved blame. This does not settle the matter, of course: perhaps a significant proportion of subjects understand "deserves blame" differently. But we currently lack data suggesting that this accounts for much of the variations in judgment listed above.

Suppose that we could get around equivocation worries, and collected a set of answers from a large group of subjects with a significant majority coming down on one side of the issue (thus satisfying conditions (4) and (5) above). Suppose further that our antecedent reasons do not very strongly favor either compatibilism or incompatibilism (thus satisfying condition (1)). Then the extent to which the answers gives us reasons to accept the majority answer depends on the *reliability* of the subjects, and the extent to which their judgments track the truth *independently*. I discuss issues of reliability in §§3–5 and issues of independence in §6.

3. Folk reliability, disagreement, and identifiable bias

Why should we think that experimental subjects are more likely than not to correctly assess whether agents in deterministic scenarios are morally responsible? A quick answer would be that even non-specialists have enough conceptual competence to form beliefs about responsibility and, if the studies are set up right, some understanding of the deterministic aspects of the scenarios they consider.

This answer might seem as misguided as it is quick. Human beings are famously subject to various perceptual and cognitive illusions, and there are issues that are simply beyond most people's ken. Based on this, one might think that reliability has to be established for the judgments at hand, and that this requires independently answering the very question that folk judgments were supposed to help us with (cf. Ludwig 2017: 395–6). But as a general demand, this is too strong. Before knowing the answer to a question, we can often determine that people have enough cognitive capacities to ground a presumption of positive reliability: pending specific reason to the contrary, reasons having to do with the matter at hand and the judge's access to it, we should expect someone with such capacities to get things right more often than not.

The question, then, is whether there are specific reasons to reject such a presumption of reliability with respect to compatibilist or incompatibilist judgments. One reason comes from disagreement about the issue at hand. For example, if 57% of a large number of judgments agreed with compatibilism and 43% disagrees with it and the judgments are independently tracking the truth, the average reliability of these judgements is unlikely to much exceed .57.[7] Given that most studies of folk incompatibilism reveal a significant minority, this might in itself restrict what degree of reliability we can plausibly attribute to subjects given some modest amount of independence. But this does not itself prevent the totality of judgments from strongly supporting one answer. If the distribution of answers is compatible with a with very modest reliability, the number of subjects is large enough, and each tracks the truth independently enough of the others, the upshot might still be a very high likelihood that the majority is correct. (As noted above, if 605 of 1,000 people who independently track truth with a meagre .505 reliability agreed on an answer, this would make that answer 99% likely to be correct!)

There are, however, more direct reasons for doubting the reliability of folk judgments about incompatibilism. The first is that the issue at hand is complex and difficult; the other is the frequent missteps displayed when non-philosophers begin to think about these issues, and the possibility that there are widespread sources that erroneously bias people's judgments in a way that swamps any general reliability. In section 5, we consider whether the nature of the issue gives us reason reject folk reliability; in this and the next section, we look closer at evidence for specific sources of bias and error.

One sign that non-philosophers are unreliable judges of incompatibilism is provided by the confusion they display when first confronted with the issue. As those who have taught classes on incompatibilism know, students not only find many of these issues difficult, but also frequently

[7] For discussion of how the distribution of answers might affect plausible attributions of reliability, see Talbot 2014: 3867–71.

make specific basic mistakes, conflating determinism and fatalism or conflating the question about whether people who do bad things can deserve blame under determinism with the question of whether there might still be reasons to blame them. The point here is not that incompatibilist judgments are mistaken, nor that compatibilist judgments are—we will soon discuss such specific claims—but that non-experts often make what are uncontroversially mistakes about matters that seem crucial for settling the question. Based on this it might seem very reasonable not to expect their acceptance or rejection of incompatibilism to be reliably truth-tracking.

The fact that non-specialists make these kinds of errors plausibly lowers their degree of reliability. But it does not necessarily mean that they will not do better than chance. As long as these various mistakes constitute noise—as long as their effects are randomly distributed—and as long as we could expect people's judgments to have a slight tendency to track the truth, and independently so, a sufficiently large majority of subjects coming down on one side might still give us strong reasons to accept that side.

Things change, however, if we know that large proportions of people are swayed by biases that prompt judgments falling specifically on one side, biases that are unlikely to be cancelled out. After discounting such judgments, the wisdom of the remaining crowd might point in the opposite direction. Here, experimental philosophers have proposed and provided empirical support for a number of hypothesis about factors that (i) drive folk judgments towards one kind of judgment and (ii) seem clearly irrelevant to the truth of the matter. In what follows, I outline and discuss three of the most important proposals.[8]

The first hypothesis is that many who make compatibilist judgments are willing to attribute responsibility *no matter what*, even to agents who are definitely not responsible. Specifically, Adam Feltz and Melissa Millan (2013) suggest, based on experimental data, that

> *No Matter What*: Many who attribute responsibility to agents in deterministic scenarios also attribute responsibility to agents who should not be judged responsible because their actions are *fated*.

An action is "fated" in the relevant sense if it is bound to happen independently of the sorts of states—character, values, preferences—that compatibilists see as grounding responsibility. Fatalism is thus different from determinism: even if our agential states are determined by factors in the distant past, our actions are still counterfactually or causally dependent on these states. Through a series of experiments, Feltz and Millan showed that those who were willing to attribute free will and responsibility to agents in various "fatalistic" scenarios (between 30% and 60% of subjects, depending on experiment) were also much more willing to do so to agents in a deterministic scenario. (Most fatalistic scenarios involved a book where the agent's decisions and actions are written beforehand and where everything written in the book will necessarily

[8] I set aside the suggestion that people's thinking about free will and moral responsibility rely on (potentially suspect) ideas about a soul (a suggestion largely debunked in Mele 2014; Vonasch et al. 2018) and the idea that judgments relied upon in manipulation arguments for incompatibilism or responsibility (or accountability) nihilism should be interpreted in compatibilist-friendly ways (proposed by Sripada 2012; debunked in Björnsson 2016).

take place.) On the assumption that fatalism does undermine free will and responsibility, this would indeed suggest that a considerable proportion of compatibilist judgments are made by unreliable judges.[9] If we could discount these judgments while taking the remaining folk judgments to be at least somewhat reliable, the overall pattern of judgments might strongly support incompatibilism.

Unfortunately for the *No Matter What* hypothesis, the kinds of fatalism that Feltz and Millan tested reactions to are not uncontroversially incompatible with responsibility. First, some of the descriptions of fated actions offered in their studies might well have been understood along the lines of determinism. Second, all their descriptions seem compatible with responsibility given popular "sourcehood" forms of compatibilism: according to these forms, inspired by Harry Frankfurt (1969), responsibility for an action only requires that it resulted from the right aspects of the agent, not that the agent could have avoided performing the action. Third, a study by James Andow and Florian Cova (2016) strongly suggests that most subjects who attribute responsibility to an agent in a fatalistic scenario take the agent's actions to be the result of such agential aspects. Though attributions of responsibility in deterministic and fatalistic scenarios of a certain kind tend to go hand in hand, then, the process at work is not one that would necessarily make the judgments in question unreliable. To show that it is, we would need independent grounds for rejecting compatibilist reasoning.

A different reason to discount apparently compatibilist judgments comes from Thomas Nadelhoffer, David Rose, Wesley Buckwalter, and Shaun Nichols (2019). They have suggested that:

> *Indeterminist Intrusion*: Many who attribute responsibility to agents in deterministic scenarios do so because they do not fully represent these scenarios as deterministic.

In experiments supporting this hypothesis, subjects were presented with deterministic scenarios and answered a number of questions, including questions about the free will and responsibility of agents in these scenarios. They were then asked what chance there was, on a scale from 0% to 100%, that these agents would do something other than what had been stipulated to follow given the past and deterministic laws. Restricting ourselves to scenarios involving human agents, well over half the subjects said that there was some non-zero chance that this would happen, apparently thinking of the scenario as involving indeterministic elements. Moreover, whereas 63% of these subjects agreed or strongly agreed with the claim that the agents were fully morally responsible for their actions, only 46% did so among those who said that there was zero chance

[9] By what processes would people erroneously attribute responsibility to fated agents? One hypothesis would be that subjects are in the grip of retributive emotions (cf. Nichols and Knobe 2007: 671–2). However, hypotheses appealing to emotional effects seem to be in conflict with the available data, judging by a recent meta-study (Feltz and Cova 2014). But there are other possibilities, not appealing to affect. One suggestion has been that we are operating with a general assumption that if something bad happens, someone is responsible (Mandelbaum and Ripley 2012). Another is that we want to believe that people are responsible, and adjust our other commitments accordingly in acts of motivated reasoning (Clark et al. 2019).

that the agent would act differently.[10] Apparently, a significant proportion of attributions of responsibility to agents in these scenarios were made without full acceptance of their key deterministic aspects. These attributions cannot plausibly increase the likelihood of compatibilism, however reliable these judges are in other regards.[11]

The *No Matter What* and *Indeterminist Intrusion* hypotheses targeted compatibilist judgments. Another recent proposal, by Eddy Nahmias and Dylan Murray (2010; 2014; cf. Nelkin 2007: 255–56), instead targets incompatibilist judgments. Nahmias and Murray suggest that

> *Bypassing*: A large proportion of people who think that agents in deterministic scenarios lack responsibility do so because they take it that these agents' rational agency is bypassed, playing no role in determining what they do.

Since determinism does not in fact imply that rational agency is bypassed, it follows from *Bypassing* that a large proportion of incompatibilist judgments are based on a clearly identifiable mistake.

The *Bypassing* hypothesis could potentially explain the *Abstract vs Concrete*, the *Neurology vs Psychology*, the *Actual vs Possible*, and the *Framing of Determinism* effects mentioned above. Cases that concern the actual world and are framed in concrete psychological terms might naturally recruit ordinary psychological explanatory models in which the effects of antecedent events on actions passes through rather than bypasses rational agency, thus counteracting the mistake. In addition, Nahmias and Murray presented data showing that in experiments varying the ways determinism was framed, incompatibilist judgments were quite strongly correlated with the acceptance of claims like

> BYPASS: In [deterministic scenario], [what a person wants / what they believe / their decisions] have no effect on what they end up doing.

Given this, it is natural to assume that variations in ways of presenting determinism give rise to varying degrees of incompatibilist judgments by prompting the bypass mistake to corresponding degrees.

---

[10] Number of subjects were 282 and 426, respectively. (Based on data set made public by the authors.) Responses could be "strongly disagree", "disagree", "somewhat disagree", "neither agree nor disagree", "somewhat agree", "agree", and "strongly agree". Studies of folk compatibilism standardly exclude subjects failing comprehension tests for determinism, but questions used for these tests tend to involve close to verbatim repetition of original description of deterministic scenario; the test used by Nadelhoffer at al. arguably required reasoning from deterministic premises.

[11] Why do people make this mistake? Interestingly, subjects thinking about deterministic scenarios involving robots rather than human agents were less inclined to make this mistake. Based on statistical analysis of this and subjects' attributions of free will to human agents and robots, Nadelhoffer et al. suggest that subjects make the error because indeterminism is part of their conception of human agency. This is in line with evidence that indeterminism is an integral part of our understanding of human conscious agency (see e.g. Deery et al. 2013; Björnsson and Shepherd 2020).

Though both striking and replicated in several later experiments, the correlation between incompatibilist and BYPASS judgments does not in the end support the *Bypassing* hypothesis.[12] First, acceptance of BYPASS statements does not well explain the acceptance of incompatibilism: the data turns out to be better accounted for by statistical models taking incompatibilist judgments to explain BYPASS judgments rather the other way around, or taking both to have a common cause (Rose and Nichols 2013; Björnsson 2014; Björnsson and Pereboom 2014). This undermines the claim that incompatibilist judgments are based on mistaken BYPASS judgments. Second, later experiments (Björnsson 2014) showed that there was no negative correlation between incompatibilist judgments and the acceptance of claims like

> THROUGHPASS: In [deterministic scenario], when earlier events cause an agent's action, they typically do so by affecting what the agent believes and wants, which in turn causes the agent to decide and act in a certain way.

In fact, there was a weak positive correlation between the acceptance of BYPASS and THROUGHPASS statements. This undermines the idea that subjects who make incompatibilist judgments really think that rational agency is bypassed.[13]

But why, then, do these subjects accept BYPASS statements? My own suggestion has been that they are gripped by the (arguably correct) thought that determinism implies that the person's desires, beliefs, and decisions, or more generally their deliberation, has no *independent* effect on their actions, being itself the mere consequence of earlier events:

> *No Independent Effect*: Subjects who (a) take determined agents' deliberation to have no independent effect on their actions and (b) take this to undermine responsibility tend to (c) interpret BYPASS statements as denying such independent causal influence because such a denial strikes them as particularly interesting (Björnsson 2014; Björnsson and Pereboom 2014).

Since BYPASS statements are (arguably) true under this interpretation, the *No Independent Effect* hypothesis gives us no reason to reject incompatibilist judgments as confused.

Another suggestion, by David Rose and Shaun Nichols, is in line with the intuitive thought that someone whose deliberation is predetermined to result in a specific intention *did not really have a choice* and *did not really decide* the already predetermined outcome:

---

[12] Even if correct, *Bypassing* needs to be reconciled with the fact that many subjects seem to be impressed by concerns of determinism when these are distinguished from concerns of conscious voluntary control. See Shepherd 2017.

[13] It also undermines Fischer's (2013) suggestion that people make incompatibilist judgments because they confusedly take determinism to undermine *guidance control* when it only threatens *regulative control.* (The former sort of control requires the right involvement in the actual causal sequence leading to the object of responsibility; the latter that one decides the unfolding of an open future.) Given this suggestion, and given that those who accept THROUGHPASS likely think that the agent has guidance control, one would expect a significant negative correlation between incompatibilist judgments and the acceptance of THROUGHPASS. The data instead reveal a weak positive correlation, and the acceptance of THROUGHPASS among a large majority of those making incompatibilist judgments (Björnsson 2014).

> *No Choice*: People who accept the BYPASS statements that determined agents' decisions do not affect their actions do so because they take determinism to imply that agents make no choices or decisions, which they in turn take to imply that they are not morally responsible for what they do (Rose and Nichols 2013).

In support of the *No Choice* hypothesis, Rose and Nichols' present data suggesting that subjects' rejection of the former of the following claims can explain their acceptance of the latter:

> DECISIONS: In this [deterministic] universe, people make decisions.
>
> NO EFFECT: In this [deterministic] universe, when people make decisions, what they think and want has no effect on what actions they end up performing.

Potentially, this could ground another error theory for incompatibilist intuitions, given the assumption that there can be such a thing as predetermined decisions.

The *No Choice* explanation is problematic, however. First, most BYPASS statements in the studies cited above merely deny the effects of what people *believe* or *desire*, and do not even mention their decisions. (Even NO EFFECT, though concerned with cases involving decisions, is denying effects of what people "think and want", not of their decisions.) And while there is something intuitive about denying that deterministic agents really have a choice, the thought that they lack beliefs or desires is much less compelling.[14]

Second, subjects who make incompatibilist judgments are no less willing to accept THROUGHPASS than those making compatibilist judgments. This suggest that they generally and easily recognize a sense in which determined agents do make decisions based on what they believe and want. This is not in itself a problem for the *No Choice* hypothesis. But if they recognize such a sense, then they presumably take NO EFFECT to be concerned with cases involving beliefs, desires, and decisions in *that* sense. The alternative would be that they understand NO EFFECT as absurdly concerned with beliefs, desires, and decisions that they think cannot exist in the deterministic universe. (So understood, NO EFFECT would be either vacuously true or suffer from a presupposition failure, being akin to "In 2018, when people met Adolf Hitler, the questions they asked had no effect on what answers he gave.") Subjects would presumably prefer an easily available interpretation not suffering from those defects. But then their rejection of DECISIONS fails to explain why they think that beliefs and desires have no effects in those cases. By contrast, the *No Independent Effect* hypothesis readily explains why incompatibilists tend to accept NO EFFECT as well as BYPASS.[15]

---

[14] Chan et al. 2016 ask bypass questions about beliefs and values separately from questions about decisions. Unfortunately, they only provide composite result. Moreover, the fairly weak correlational data presented in the article make it hard to see how bypass judgments can be largely explained by responsibility judgments.

[15] It can also be extended to explain why the rejection of DECISION is correlated with the acceptance of NO EFFECT and the rejection of responsibility. Assume, as seems likely, that to "make a decision" can be understood strongly, as requiring that one plays a causally independent role in forming an otherwise undetermined intention, as well as weakly, as merely requiring that one's deliberation figures in a certain way in the causal sequence leading to that intention. Just as with the "no independent effect" interpretation of the BYPASS and NO EFFECT statements, the stronger interpretation will be more

The *No Matter What*, *Indeterminist Intrusion*, and *Bypassing* hypotheses represent what I take to be the most promising recent attempts to experimentally establish that a large proportion of either compatibilist or incompatibilist verdicts are epistemically defective. *No Matter What* failed to identify a process of judgment formation that is uncontroversially epistemically suspect. *Bypassing* identified such a process, but data contradicted the claim that incompatibilist judgments result from it. *Indeterminist Intrusion*, finally, identified a process rendering some attributions of responsibility to determined agents suspect, though a large proportion of judgments not suspect in this way were still compatibilist judgments.

4. Reliability and conditional likelihoods

Though not discussed in the literature, there is another way in which hypotheses about epistemic processes might be important, one that does not rely on these processes being *uncontroversially suspect*. Reliability in judgment, remember, is understood as the likelihood of making a correct judgment: of judging that P if P is true and judging that not-P if not-P is true. For a claim about an epistemic process to affect the relevant sort of reliability, it is thus enough to make plausible that the process would affect the likelihood of a certain kind of judgment *given that a certain answer is true*. This makes it easier to avoid begging the question against one answer. Nevertheless, conditional likelihoods can be epistemically highly relevant. Suppose, for example, that we have a plausible account of why many people would reject compatibilism even if it were true, but no plausible account of why many people would reject incompatibilism if it were true. Then the likelihood of a compatibilist judgment given that compatibilism is true might now be .5 whereas the likelihood of an incompatibilist judgment given incompatibilism might be .75. Suppose also that the antecedent likelihood of incompatibilism is .75, that judges are tracking truth independently, and that 58 out of 100 judgments are incompatibilist. Then in spite of the preponderance of incompatibilist judgments and the above chance antecedent likelihood that an incompatibilist judgment will be correct, the *compatibilist* position would be .99 likely to be correct! This might look surprising, but the explanation is simple: the distribution of answers would be significantly more likely given compatibilism than given incompatibilism, involving fewer and less unlikely independent errors.

Conditional likelihoods can in principle provide non-question begging evidence for a position. But any defense of such likelihoods will also tend to be much more complex than the sorts of arguments we have seen for the *No Matter What*, *Indeterminist Intrusion*, and *Bypassing* hypotheses. To assess the likelihood of getting things right given that a certain answer to the question is correct might require (i) a sense of what the relation of responsibility is if that answer is correct, (ii) an explanation of what makes *that* the relation that need to obtain for our responsibility attributions to be true, and thus (iii) a plausible general account of what determines the truth-conditions of classes of judgments, as well as (iv) an account of how our

interesting and thus tend to be more salient for subjects who take determinism to rule out independent causal influence and take this to undermine responsibility.

judgments might track or fail to track these facts. While experimental philosophy can have a role to play here, it would be in the context of systematic philosophical work.[16]

5. Folk reliability, difficult philosophical questions, and conceptual analysis

Suppose that we have no particular evidence of systematic mistakes that would make people no better than chance in their responsibility judgments about deterministic scenarios (setting aside those who fail to grasp key aspects of determinism). Suppose also that we lack evidence that compatibilist judgments are more likely to be mistaken, or that incompatibilist judgments are. We might still conclude that non-specialists do no better than chance with this particular question because of its difficulty and complexity. Generally, the issue seems no different from other central philosophical topics, with its striking yet hard to pin down relations to non-trivial matters such as causation, explanation, laws, control, freedom, rationality, agency, moral wrongness, blame, credit, reactive attitudes, punishment, and desert. Moreover, non-specialists lack access to the kinds of arguments that philosophers take to be importantly revealing, such as Peter van Inwagen's (1983) consequence argument, Harry Frankfurt's (1969) argument against the requirement of alternative possibilities, or Peter Strawson's (1962) account of the systematic similarities between personal reactive attitudes and practices of holding responsible, to mention some of the most influential (cf. Sommers 2010: 205–6). Given this, how could folk judgments be more reliable than chance?

Perhaps, though, this is the wrong way to think about the subject matter. On one understanding, judgments about imagined deterministic cases are based on direct, intuitive, knowledge of relatively simple necessary truths about responsibility and an understanding of basic features of determinism, rather that knowledge about complex ethical, metaphysical, or empirical matters. Because confusion of the latter factors and an insecure grasp of the relevant concepts downgrades intuitive access to the relevant facts, philosophical arguments can help us see more clearly, but they are not strictly speaking necessary. (Contrast knowledge of whether incompatibilism is true with knowledge of whether human decision making is a deterministic process. The latter is clearly an empirical matter to which non-specialists have no independent access.[17] Incompatibilism does not seem to be inaccessible in *that* way.) On one popular version of this view, the facts are conceptual truths, and they are accessed by the employment of the concepts involved, concepts possessed by non-specialists and specialists alike.[18] In either case,

[16] For my own attempt to explain why, if compatibilism is correct, it would be true in a way given which we should expect the messy distribution of judgments canvassed in section 2, see Björnsson and Persson 2012; 2013; Björnsson 2014; 2016; 2017. For other systematic attempts to incorporate epistemic and semantic reflection in theorizing about free will and moral responsibility, see Double 1996; McCormick 2013; Vargas 2013; Nichols 2015; Deery 2019; Kumar forthcoming.

[17] It is of course also an issue not yet resolved by science (see e.g. Balaguer 2010).

[18] For a defense of the potential usefulness of experimental philosophy in revealing conceptual truths, see Balaguer 2016.

non-specialists can access the relevant facts, even if their susceptibility to various sources of noise make their reliability low.[19]

Many metasemantic accounts—accounts of how words or elements of thought have their referents determined—reject the idea that we have privileged access to the nature or necessary features of these referents merely in virtue of our rational capacities and possession of the relevant concepts.[20] But even on such accounts, it might seem plausible that people in general can have access to the truth conditions of responsibility judgments. For a recent example, consider Laura and François Schroeter's (2014) explanation of how people can be concerned with the very same subject matter even if they have very different conceptions of it. The key, they suggest, is that the parties take themselves to be concerned with what they and others have been concerned with at other times, and thus as continuing one and the same "representational tradition". What property this tradition has been concerned with, if any, is then a matter of what interpretation makes best sense of the tradition as a whole, not about individual judgments based on different conceptions of the subject matter. Conflicting conceptions might in turn be sustained by conflicting attempts to make sense of these traditions, fueled when traditions involve tensions that might be resolved in different ways depending on which of their commitments are seen as more central.[21] At least judging from the philosophical literature, this is very much true about the tradition of attributing responsibility in the context of guiding reactive attitudes and practices of holding responsible, with its notorious tension between compatibilist and incompatibilist or skeptical elements.[22] But even with all this variation, it seems that individual judges, qua members of the relevant tradition, will tend to have access to the very aspects of it that a best interpretation should make sense of. It would be odd to antecedently deny that this access makes members better than chance at getting right whether determinism would rule out responsibility.[23]

---

[19] It is also unclear to what extent experts do better, and with regard to what judgments. For an introductory discussion, see e.g. Nado 2014.

[20] Classical discussions of how our most fundamental ideas about a thing can come apart from what that idea is about include Putnam 1973; 1975; Kripke 1980; Lewis 1984; Millikan 2000.

[21] As the Schroeters stress (2014: 16–18), the best interpretation will at times require disambiguation. Moreover, as Mark Balaguer insists (2016: 2379–80), we should allow that people might have inconsistent and thus uninstantiated concepts in mind. But this does not itself mean that every inconsistency in dispositions to apply a concept calls for disambiguation or means that the concept to which judges relate is itself inconsistent.

[22] Cf. Vargas 2006; 2013, part I. A possibility here is that different ways of resolving tensions might be more or less appropriate depending on context, leaving room for even more variations in judgment (cf. Nichols 2015).

[23] Not all metasemantic theories are neutral with respect to compatibilism and incompatibilism. For example, consider teleosemantic accounts like that of Ruth Millikan's (1984; 2017), which take the referent of a concept to be a matter of what the mechanisms involved in employing that concept have as their etiological function to recognize. For mechanisms that produce judgments that regularly guide action and communication in what people take to be successful ways, that function is not plausibly to track some metaphysically esoteric condition without systematic effects on the practice in which those mechanisms are employed. Cf. Björnsson and Persson 2012: 345–46; Vargas 2013, ch. 3; Deery 2019.

6. Are people independently tracking the truth?

Suppose that we can accord people some reliability—perhaps a very modest one. Suppose further that considerably more judgments are in line with one theory than with its rival. Then these judgments could provide very strong evidence for that theory in the way sketched in section 1. To do so, however, they must be independently tracking the truth of the matter, in the following sense:

> *Independence*: One judgment whether P is (completely) independent of other judgments whether P if and only if (i) the likelihood that it will yield the verdict that P on the assumption that P and (ii) the likelihood that it will yield the verdict that not-P on the assumption that not-P are unaffected by what we assume that the verdicts of the other judgments are.

Independence, so understood, would most obviously be undermined if we knew that some judgments would causally influence others, perhaps because some judges took others to be authorities on the issue. For an extreme example, suppose that we knew that only one judge would make a judgment based on inconclusive evidence and everyone else would merely try to copy that judgment (with fairly high reliability). Then the likelihood that one particular copycat judgment, J, would yield the verdict that P on the assumption that P would depend on what we assume that the other copycat judgments are. For these judgments, especially in aggregate, would tell us something about the original judgment from which it would be copied and about the evidence on which the original judgment was based. Thus, if we assume that P and that all the other copycats judged that P, it would be very likely that J would yield the verdict that P, but if we assumed that P and that all the other copycats judged that not-P, it would be very likely that J would yield the verdict that not-P. Moreover, it is clear that even if all judgments agreed and the original judgment was reasonable in light of the evidence, the aggregate reliability of all judgments would be no higher than the reliability of the evidence, and the aggregate reliability of the copycat judgments would be no higher than the reliability of the original judgment.[24]

Admittedly, causal influence of some judgments on others might not be an urgent worry for the sorts of studies that we have looked at. Subjects are often recruited from large online pools rather than from small socially connected groups, and even when subjects are socially connected, the sorts of judgments reported in the experiments are not characteristically socially shared. But it seems highly likely that many judges have access to the same kind of evidence. Many might rely on intuitive reactions to the cases that confront them in the studies prompted by similar cognitive processes, and many will have similar experience of holding and being held responsible, and of various appeals to excuses and exemptions. Even if there is a richness and variation in the evidence on which people base their judgments, there is bound to be a great deal of overlap. It could be, then, that although more judgments will tend to bring higher

[24] For discussion of the independence requirement, see e.g. Estlund 1994; Dietrich 2008; Talbot 2014; Goodin and Spiekermann 2018, ch. 5.

accumulated reliability, the marginal epistemic utility might fall quickly, as more judges would be increasingly unlikely to bring new sources of reliability to the table rather than more of the same. In light of this, and in light of the fact that professional philosophers have access to evidence in addition to what they share with laypeople, the wisdom of large crowds of non-specialists might add little to the judgments of specialists.

7. Concluding discussion

Depending on the subject matter, a preponderance of judgments in favor of one answer might mean that it is extremely likely to be correct. We have seen several reasons to doubt that it applies to the central philosophical question discussed here, namely that of compatibilism. We can perhaps assume, as most people in the literature do, that the issue of whether compatibilism is true has one correct answer, and might well accord people some antecedent reliability. Pending further arguments, however, the reliability is presumably low, as the questions are difficult and people seem to be fairly evenly divided in their judgments. More importantly, it is unclear to what extent we should assume that people's judgments are independently tracking the truth. If they are not, their individual reliabilities fail to aggregate.

This mostly negative conclusion is admittedly concerned with one specific and notoriously thorny philosophical issue, but it is easy to see how it generalizes to other similarly difficult questions. Much of what is said also applies if we turn from appeals to the wisdom of crowds of non-specialists to corresponding appeals to the judgments of specialists. Going by the raw numbers of compatibilists among specialists in action theory (see Bourget and Chalmers 2009; 2014) and assuming a modest reliability and independence and an antecedently equal likelihood for compatibilism and incompatibilism, compatibilism would be highly likely to be correct, whereas libertarianism and free will skepticism would be extremely unlikely.[25] One can resist this conclusion by pessimistically denying that specialists are even modestly reliable. But, less pessimistically, one can deny the independence of these judgments. As with laypeople, there are good reasons to think that philosophers' reliability depends on their access to largely overlapping sources of evidence mostly falling into a few large groups whose reliability are strongly interdependent.[26]

The worries raised here have concerned the idea that when a preponderance of judgments by non-specialists fall in line with one philosophical position, this can provide strong aggregative evidence for that position. But studies of folk judgments might help us understand responsibility in less direct ways. First, they might serve as a check on philosophical group think. If one suspects that the field ignores an important strand of ordinary thinking about responsibility, experiments involving non-philosophers can tell us whether it is indeed an important part of the practice that needs to be accounted for or explained away. Second, although the *No Matter What*

[25] These raw numbers might well be misleading, as subjects were asked about free will rather than moral responsibility, and as some who call themselves compatibilists about free will have in mind a kind of responsibility that few hard determinists or responsibility skeptics deny.

[26] For discussion of how dependence undermines aggregation of evidential weight among philosophers, see Talbot 2014.

and *Bypassing* hypotheses did not hold up in further studies, *Indeterminist Intrusion* currently seems well supported, and new experiments might reveal further reasons why certain judgments are particularly unreliable and ways in which compatibilists or incompatibilists can explain away judgments that do not accord with their theory. (Studies of philosopher's judgments could reveal the same, calling for methodological self-reflection.) Third, studies might reveal widespread dispositions of judgment and relations between different kinds of judgments that tell us something about the nature of the practice.[27] But, as noted in the introduction, neither of these ways for experimental philosophy to be helpful bypasses the need to attend to more ordinary philosophical arguments and systematizing efforts. Though it can be a useful tool, it is no shortcut.

Acknowledgments

I thank Manuel Vargas, Christopher Franklin, Shaun Nichols, Mark Balaguer, Martin Peterson, Kai Spiekermann, Dana Nelkin, and Derk Pereboom for valuable input on various drafts. Work on this paper was funded by the Swedish Research Council [2015-01488].

Bibliography

Alexander, Joshua, Mallon, Ronald and Weinberg, Jonathan 2010: 'Accentuate the Negative'. *Review of Philosophy and Psychology*, 1, pp. 297-314.

Andow, James and Cova, Florian 2016: 'Why Compatibilist Intuitions Are Not Mistaken: A Reply to Feltz and Millan'. *Philosophical Psychology*, 29, pp. 550-66.

Balaguer, Mark 2010: *Free Will as an Open Scientific Problem*. Cambridge, MA: The MIT Press.

Balaguer, Mark 2016: 'Conceptual Analysis and X-Phi'. *Synthese*, 193, pp. 2367-88.

Björnsson, Gunnar 2014: 'Incompatibilism and 'Bypassed' Agency'. In *Surrounding Free Will*. Mele, Alfred (ed) New York: Oxford University Press pp. 95–122.

Björnsson, Gunnar 2016: 'Outsourcing the Deep Self: Deep Self Discordance Does Not Explain Away Intuitions in Manipulation Arguments'. *Philosophical Psychology*, 29, pp. 637-53.

Björnsson, Gunnar 2017: 'Explaining Away Epistemic Skepticism About Culpability'. In *Oxford Studies in Agency and Responsibility*. Shoemaker, David (ed) Oxford University Press pp. 141–64.

Björnsson, Gunnar and Pereboom, Derk 2014: 'Free Will Skepticism and Bypassing'. In *Moral Psychology: Free Will and Moral Responsibility*. Sinnott-Armstrong, Walter (ed) MIT Press pp. 27–35.

Björnsson, Gunnar and Pereboom, Derk 2016: 'Traditional and Experimental Approaches to Free Will and Moral Responsibility'. In *A Companion to Experimental Philosophy*. Wiley pp. 142-57.

Björnsson, Gunnar and Persson, Karl 2012: 'The Explanatory Component of Moral Responsibility'. *Noûs*, 46, pp. 326–54.

Björnsson, Gunnar and Persson, Karl 2013: 'A Unified Empirical Account of Responsibility Judgments'. *Philosophy and Phenomenological Research*, 87, pp. 611–39.

Björnsson, Gunnar and Shepherd, Joshua 2020: 'Determinism and Attributions of Consciousness'. *Philosophical Psychology*, 33, pp. 549-68.

Bourget, David and Chalmers, David J. 2009: 'The Philpapers Survey'. p. https://philpapers.org/surveys/.

[27] See e.g. (a) Victor Kumar's (forthcoming) account of responsibility and associated retributive practices, based in part on Cushman 2008, (b) Karl Persson's and my appeal to empirical studies in support of a certain account of the nature of responsibility judgments (Björnsson and Persson 2012; 2013), and (c) Shaun Nichol's (2015) account of retributive practices and deterministic and incompatibilist intuitions.

Bourget, David and Chalmers, David J. 2014: 'What Do Philosophers Believe?'. *Philosophical Studies*, 170, pp. 465-500.
Chan, Hoi-Yee, Deutsch, Max and Nichols, Shaun 2016: 'Free Will and Experimental Philosophy'. In *A Companion to Experimental Philosophy*. John Wiley & Sons. pp. 158–72.
Clark, Cory J., Winegard, Bo M. and Baumeister, Roy F. 2019: 'Forget the Folk: Moral Responsibility Preservation Motives and Other Conditions for Compatibilism'. *Frontiers in Psychology*, 10.
Cova, Florian, Strickland, Brent, Abatista, Angela, Allard, Aurélien, Andow, James, Attie, Mario, Beebe, James, Berniūnas, Renatas, Boudesseul, Jordane, Colombo, Matteo, Cushman, Fiery, Diaz, Rodrigo, N'Djaye Nikolai van Dongen, Noah, Dranseika, Vilius, Earp, Brian, Torres, Antonio Gaitán, Hannikainen, Ivar, Hernández-Conde, José V., Hu, Wenjia, Jaquet, François, Khalifa, Kareem, Kim, Hanna, Kneer, Markus, Knobe, Joshua, Kurthy, Miklos, Lantian, Anthony, Liao, Shen-yi, Machery, Edouard, Moerenhout, Tania, Mott, Christian, Phelan, Mark, Phillips, Jonathan, Rambharose, Navin, Reuter, Kevin, Romero, Felipe, Sousa, Paulo, Sprenger, Jan, Thalabard, Emile, Tobia, Kevin, Viciana, Hugo, Wilkenfeld, Daniel and Zhou, Xiang 2018: 'Estimating the Reproducibility of Experimental Philosophy'. *Review of Philosophy and Psychology*.
Cushman, Fiery 2008: 'Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment'. *Cognition*, 108, pp. 353-80.
De Brigard, Felipe, Mandelbaum, Eric and Ripley, David 2009: 'Responsibility and the Brain Sciences'. *Ethical Theory and Moral Practice*, 12, pp. 511-24.
Deery, Oisín 2019: 'Free Actions as a Natural Kind'. *Synthese*.
Deery, Oisín, Bedke, Matthew S. and Nichols, Shaun 2013: 'Phenomenal Abilities: Incompatibilism and the Experience of Agency'. In *Oxford Studies in Agency and Responsibility*. Shoemaker, David (ed) Oxford UP pp. 126–50.
Dietrich, Franz 2008: 'The Premises of Condorcet's Jury Theorem Are Not Simultaneously Justified'. *Episteme*, 5, pp. 56-73.
Double, Richard 1996: *Metaphilosophy and Free Will*. Oxford University Press.
Estlund, David M. 1994: 'Opinion Leaders, Independence, and Condorcet's Jury Theorem'. *Theory and Decision*, 36, pp. 131-62.
Feltz, Adam and Cova, Florian 2014: 'Moral Responsibility and Free Will: A Meta-Analysis'. *Consciousness and Cognition*, 30, pp. 234-46.
Feltz, Adam and Millan, Melissa 2013: 'An Error Theory for Compatibilist Intuitions'. *Philosophical Psychology*, pp. 1-27.
Fischer, John Martin 2013: 'The Frankfurt Style Cases: Philosophical Lightning Rods'. In *In Free Will and Moral Responsibility*. Haji, Ish and Caouette, Justin (eds) Newcastle: Cambridge Scholars Publishing pp. 43–57.
Frankfurt, Harry G. 1969: 'Alternate Possibilities and Moral Responsibility'. *Journal of Philosophy*, 66, pp. 829-39.
Goodin, Robert E and Spiekermann, Kai 2018: *An Epistemic Theory of Democracy*. Oxford University Press.
Jeppsson, Sofia 2016: 'Accountability, Answerability, and Freedom'. *Social Theory & Practice*, 42, pp. 681-705.
Kauppinen, Antti 2007: 'The Rise and Fall of Experimental Philosophy'. *Philosophical Explorations*, 10, pp. 95-118.
Kneer, Markus and Machery, Edouard 2019: 'No Luck for Moral Luck'. *Cognition*, 182, pp. 331-48.
Knobe, Joshua 2007: 'Experimental Philosophy and Philosophical Significance'. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, 10, pp. 119 - 21.
Knobe, Joshua and Nichols, Shaun 2017: 'Experimental Philosophy'. In *The Stanford Encyclopedia of Philosophy* Zalta, Edward N. (ed) URL= https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/.
Kripke, Saul A. 1980: *Naming and Necessity*. Harvard University Press.
Kumar, Victor forthcoming: 'Empirical Vindication of Moral Luck'. *Noûs*.
Lewis, David 1984: 'Putnam's Paradox'. *Australasian Journal of Philosophy*, 62, pp. 221-36.
List, Christian 2004: 'On the Significance of the Absolute Margin'. *British Journal for the Philosophy of Science*, 55, pp. 521-44.

Ludwig, Kirk 2017: 'Thought Experiments in Experimental Philosophy'. In *Routledge Companion to Thought Experiments*. Stuart, Mike, Brown, James Robert and Fehige, Yiftach J. H. (eds) Routledge pp. 385-405.
Mallon, Ron 2016: 'Experimental Philosophy'. *Oxford handbook of philosophical methodology*, pp. 410-33.
Mandelbaum, Eric and Ripley, David 2012: 'Explaining the Abstract/Concrete Paradoxes in Moral Psychology: The Nbar Hypothesis'. *Review of Philosophy and Psychology*, 3, pp. 351-68.
McCormick, Kelly Anne 2013: 'Anchoring a Revisionist Account of Moral Responsibility'. *Journal of Ethics and Social Philosophy*, 7, pp. 1-20.
Mele, Alfred 2014: 'Free Will and Substance Dualism: The Real Scientific Threat to Free Will?'. In *Moral Psychology, Vol. 4: Free Will and Responsibility*. Sinnott-Armstrong, Walter (ed) MIT Press pp. 195–207.
Millikan, Ruth Garrett 1984: *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.
Millikan, Ruth Garrett 2000: *On Clear and Confused Ideas: An Essay About Substance Concepts*. New York: Cambridge U. P.
Millikan, Ruth Garrett 2017: *Beyond Concepts: Unicepts, Language, and Natural Information*. Oxford University Press.
Murray, Dylan and Nahmias, Eddy 2014: 'Explaining Away Incompatibilist Intuitions'. *Philosophy and Phenomenological Research*, 88, pp. 434-67.
Nadelhoffer, Thomas, Rose, David, Buckwalter, Wesley and Nichols, Shaun 2019: 'Natural Compatibilism, Indeterminism, and Intrusive Metaphysics'. *OSF Preprints*, August 25.
Nado, Jennifer 2014: 'Philosophical Expertise'. *Philosophy Compass*, 9, pp. 631-41.
Nahmias, Eddy, Coates, D. Justin and Kvaran, Trevor 2007: 'Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions'. *Midwest Studies in Philosophy*, 31, pp. 214-42.
Nahmias, Eddy, Morris, Stephen, Nadelhoffer, Thomas and Turner, Jason 2005: 'Surveying Freedom: Folk Intuitions About Free Will and Moral Responsibility'. *Philosophical Psychology*, 18, pp. 561 - 84.
Nahmias, Eddy and Murray, Dylan 2010: 'Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions'. In *New Waves in Philosophy of Action*. Aguilar, Jesús, Buckareff, Andrei and Frankish, Keith (eds) Palgrave Macmillan pp. 189-216.
Nelkin, Dana K. 2007: 'Do We Have a Coherent Set of Intuitions About Moral Responsibility?'. *Midwest Studies in Philosophy*, 31, pp. 243-59.
Nichols, Shaun 2004: 'The Folk Psychology of Free Will: Fits and Starts'. *Mind & Language*, 19, pp. 473-502.
Nichols, Shaun 2015: *Bound: Essays on Free Will and Responsibility*. Oxford University Press.
Nichols, Shaun and Knobe, Joshua 2007: 'Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions'. *Noûs*, 41, pp. 663-85.
Paolacci, Gabriele and Chandler, Jesse 2014: 'Inside the Turk: Understanding Mechanical Turk as a Participant Pool'. *Current Directions in Psychological Science*, 23, pp. 184-88.
Putnam, Hilary 1973: 'Meaning and Reference'. *Journal of Philosophy*, 70, pp. 699-711.
Putnam, Hilary 1975: 'The Meaning of 'Meaning''. In *Mind, Language and Reality*. Cambridge: Cambridge U. P. pp. 215-71.
Rose, David and Nichols, Shaun 2013: 'The Lesson of Bypassing'. *Review of Philosophy and Psychology*, 4, pp. 599-619.
Roskies, Adina L. and Nichols, Shaun 2008: 'Bringing Responsibility Down to Earth'. *Journal of Philosophy*, 105, pp. 371–88.
Sarkissian, Hagop, Chatterjee, Amita, Brigard, Felipe De, Knobe, Joshua, Nichols, Shaun and Sirker, Smita 2010: 'Is Belief in Free Will a Cultural Universal?'. *Mind & Language*, 25, pp. 346-58.
Schroeter, Laura and Schroeter, François 2014: 'Normative Concepts: A Connectedness Model'. *Philosophers' Imprint*, 14, pp. 1–26.
Shepherd, Joshua 2015: 'Consciousness, Free Will, and Moral Responsibility: Taking the Folk Seriously'. *Philosophical Psychology*, 28, pp. 929-46.
Shepherd, Joshua 2017: 'The Folk Psychological Roots of Free Will'. In *Experimental Metaphysics*. Rose, David (ed) Bloomsbury Academic pp. 95–116.
Shoemaker, David 2015: *Responsibility from the Margins*. Oxford University Press.
Smith, Angela M. 2015: 'Responsibility as Answerability'. *Inquiry*, pp. 1-28.

Sommers, Tamler 2010: 'Experimental Philosophy and Free Will'. *Philosophy Compass*, 5, pp. 199-212.
Sripada, Chandra Sekhar 2012: 'What Makes a Manipulated Agent Unfree?'. *Philosophy and Phenomenological Research*, 85, pp. 563-93.
Strawson, Peter F. 1962: 'Freedom and Resentment'. *Proceedings of the British Academy*, 48, pp. 187–211.
Sytsma, Justin and Buckwalter, Wesley 2016: *A Companion to Experimental Philosophy*. John Wiley & Sons.
Talbot, Brian 2014: 'Why So Negative? Evidence Aggregation and Armchair Philosophy'. *Synthese*, 191, pp. 3865-96.
Taylor, Matthew C. and Maranges, Heather M. Forthcoming: 'Are the Folk Historicists About Moral Responsibility?'. *Philosophical Psychology*.
van Inwagen, Peter 1983: *An Essay on Free Will*. Oxford: Clarendon.
Vargas, Manuel 2006: 'Philosophy and the Folk: On Some Implications of Experimental Work for Philosophical Debates on Free Will'. *Journal of Cognition and Culture*, 6, pp. 239-54.
Vargas, Manuel 2013: *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.
Vonasch, Andrew J., Baumeister, Roy F. and Mele, Alfred R. 2018: 'Ordinary People Think Free Will Is a Lack of Constraint, Not the Presence of a Soul'. *Consciousness and Cognition*, 60, pp. 133-51.
Watson, Gary 1996: 'Two Faces of Responsibility'. *Philosophical Topics*, 24, pp. 227– 48.
Weinberg, Jonathan 2015: 'The Methodological Necessity of Experimental Philosophy'. *DISCIPLINE FILOSOFICHE*, 25, pp. 23–42.