Traditional and Experimental Approaches to Free Will and Moral Responsibility

Gunnar Björnsson and Derk Pereboom

The Blackwell Companion to Experimental Philosophy, Wesley Buckwalter and Justin Sytsma, eds., Oxford: Blackwell Publishers, 2016, pp. 142-57 .
Penultimate Version

1. Introduction

From the early days of experimental philosophy, attention has been focused on the problem of free will and moral responsibility. This is a natural topic for this methodology, given its proximity to the universal concerns of human life, together with the intensity with which the issues are disputed. We'll begin by introducing the problem and the standard terminology used to frame it in the philosophical context. We'll then turn to the contributions of experimental philosophy, and the prospects for the use of this methodology in the area.

The problem of free will and moral responsibility arises from a conflict between two powerful considerations. On the one hand, we human beings typically believe that we are in control of our actions in a particularly weighty sense. We express this sense of difference when we attribute moral responsibility to human beings but not, for example, to machines like thermostats and computers. Traditionally, it's supposed that moral responsibility requires us to have some type of free will in producing our actions, and hence we assume that humans, by contrast with such machines, have this sort of free will. At the same time, there are reasons for regarding human beings as relevantly more like mechanical devices than we ordinarily imagine. These reasons stem from various sources: most prominently, from scientific views that consider human beings to be components of nature and therefore governed by natural laws, and from theological concerns that require everything that occurs to be causally determined by God.

One threat to our having the sort of free will required for moral responsibility results from the view that the natural laws are deterministic, which motivates the position that all of our actions are causally determined by factors beyond our control. An action will be causally determined in this way if a process governed by the laws of nature and beginning with causally relevant factors prior to the agent's coming to be ensures the occurrence of the action. An action will also be causally determined by factors beyond the agent's control if its occurrence is ensured by a causal process that originates in God's will and ends with the action. For many contemporary philosophers, the first, naturalistic version of causal determinism about action is a serious possibility, and thus the threat that it poses to our conception of ourselves as morally responsible for our actions is serious and prevalent.

The history of philosophy records three standard types of reaction to this threat. *Compatibilists* maintain that it is possible for us to have the free will required for moral responsibility if determinism is true. Others argue that determinism is not compossible with our having the free will required for moral responsibility—they are *incompatibilists*—but they resist the

reasons for determinism and claim that we do possess this free will of this kind. They advocate the *libertarian* position. *Hard determinists* are also incompatibilists, but they contend that determinism is true and that we lack the sort of free will required for moral responsibility; they are, consequently, *free will skeptics*.

Especially since David Hume's discussion of these issues (1739/1978; 1748/2000), the concern about our having the sort of free will required for moral responsibility has been extended to whether it is compatible with the *indeterminacy* of actions. One worry is that if an action is an undetermined event, then its occurring rather than not will not be under the agent's control. According to one of the interpretations of quantum mechanics, undetermined events occur at the quantum level. One might imagine that there are structures in the brain that allow this kind of indeterminism to percolate up to the level of action, so that our actions are often undetermined. A concern for this sort of view is that agents don't control whether quantum-level undetermined events occur rather than not, and so it would seem that they would not control whether the actions to which such events give rise occur rather than not.

This development has challenged the value of the threefold classification just canvassed, despite its persistence in the contemporary debate. In particular, some maintain that the free will required for moral responsibility is not only incompatible with determinism, but in addition with at least some varieties of indeterminism. Agent-causal libertarians typically hold that this kind of free will is incompatible with the kind of indeterminism according to which only events are causes. Many free-will skeptics agree. A skeptic such as Galen Strawson maintains that this kind of free will is incompatible with any variety of indeterminism (1986; 1994). Pereboom (2001; 2014) argues that it is incompatible only with the event-causal sort, and not with indeterministic agent causation, but that the type of indeterministic agent causation that could secure moral responsibility is empirically implausible.

Complications arise on the compatibilist side as well. Hume, and later R. E. Hobart (1934) and A. J. Ayer (1954), contend that while the sort of free will required for moral responsibility is compatible with determinism, it is in fact incompatible with indeterminism, at least with indeterminism located at the point at which a decision or an intention is produced. This tradition continues in the work of philosophers such as Ishtiyaque Haji (1998) and Alfred Mele (2006). A different sort of compatibilism, according to which this sort of free will is compatible with both determinism and indeterminism, is inspired by some remarks of Hume's, and is developed in detail by P. F. Strawson in his "Freedom and Resentment" (1962). In this view, the practice of holding people morally responsible has its own internal system of norms, but is not properly subject to an external challenge, from, for example, general scientific discoveries about the universe. Whether the universe is causally deterministic or indeterministic is claimed to be irrelevant to whether our holding agents morally responsible is legitimate, and in this respect Strawson's compatibilism is *insulationist*.

It is important to recognize that our practice of holding morally responsible has a variety of aims, and that this plausibly gives rise to a number of different senses of moral responsibility. There is one particular sense of moral responsibility, together with a correlated sense of free will -- free will as the control in action required for moral responsibility in this sense -- that has been at play in the historical debate:

> For an agent to be *morally responsible for an action in the basic desert sense* is for it to belong to her in such a way that she would deserve to be blamed if she understood that it was morally wrong, and she would deserve to be credited or perhaps praised if she understood that it was morally exemplary. The desert invoked is basic in the sense that the agent, to be morally responsible, would deserve to be blamed or credited just because she performed the action, given sensitivity to its moral status, and not by virtue of consequentialist or contractualist considerations.

Basic desert moral responsibility is arguably presupposed by our retributive reactive attitudes, such as indignation and moral resentment. In P. F. Strawson's (1962) account, moral responsibility is essentially tied to these reactive attitudes, and hence the basic-desert entailing sense is plausibly the variety that he brings to the fore.

Incompatibilists hold that causal determination is incompatible with basic desert moral responsibility and with the sort of free will required for it. (Carolina Sartorio (2014) convincingly argues that causal determination of factors involving the agent, by contrast with causal determinism per se, poses the threat). However, rejecting the possibility of moral responsibility in this sense leaves other senses intact. For instance, when we encounter apparently immoral behavior, we consider it legitimate to ask the agent, "Why did you decide to do that?" or, "Do you think it was the right thing to do?" If the reasons given in response to such questions are morally unsatisfactory, we regard it as justified to invite the agent to evaluate critically what his actions indicate about his intentions and character, to demand apology, or to request reform (Scanlon 1998; Smith 2008; McKenna 2012). Engaging in such interactions is reasonable in light of the right of those harmed or threatened to protect themselves from immoral behavior and its consequences. In addition, we might have a stake in reconciliation with the wrongdoer, and calling him to account in this way can function as a step toward realizing this objective. We also have an interest in his moral formation, and the address described naturally functions as a stage in this process (Pereboom 2013, 2014; cf. Vargas 2013). The main thread of the historical free will debate does not pose causal determination as a challenge to moral responsibility conceived in this way, and free will skeptics can accept that we are morally responsible in this sense.

2. The relevance of experimental studies of responsibility judgments

Most arguments for or against the possibility of free will and moral responsibility rely on premises that the various participants in the debate would find intuitively appealing, whether these premises take the form of general principles or verdicts about particular cases. Some of these premises have clear empirical or a posteriori components: claims about the laws of nature, the existence of certain kinds of causes, the role that judgments of

responsibility play in governing our practices of holding responsible, and the effects of these practices on human beings. To assess such claims, there is clearly a role for systematic empirical investigation. However, much of the empirical work done by philosophers in this area—the sort of work that has typically been associated with experimental philosophy—has focused on the judgments of non-philosophers, in particular on whether non-philosophers have compatibilist or incompatibilist beliefs. It is natural to wonder just how studies of "folk judgments" can be relevant to the traditional philosophical questions.

Questions about what affects people's judgments of moral responsibility, or about whether or not people tend to be compatibilists, can be interesting in themselves. But given how difficult and subtle many issues of relevance to compatibilism are even for specialists, why expect help from judgments of people less trained in making relevant distinctions and assessing abstract claims, and less familiar with what has been said and argued? True, both compatibilist and incompatibilist philosophers have made claims about what ordinary people believe about free will and moral responsibility (for examples, see Nahmias et al. 2006, 29–30), and such claims are best tested empirically. What is not clear is how much stock philosophers should put in such ordinary beliefs. Accordingly, before examining some recent experimental contributions, we'll begin by canvassing some of the more prominent reasons why philosophers concerned with traditional questions about moral responsibility might take an interest in folk judgments and folk conceptions.

One reason for philosophers to care about whether their accounts conform to folk conceptions of responsibility is terminological. For instance, for these accounts to be accounts of *moral responsibility* rather than of some other relation, they had better be about what people in general have in mind when using the term 'responsibility' in the relevant moral contexts. If it turns out that the folk have nothing consistent or determinate in mind, these accounts might instead be seen as attempts to make the folk conceptions more precise. In addition, an account of the preconditions for responsibility that rejects a central part of folk conceptions of responsibility should be viewed as revisionary, and thus in need of special justification (Vargas 2013). Still, this terminological constraint is rather weak if our question concerns the preconditions for moral responsibility in the basic desert sense. In particular, it seems to matter little whether people in general associate the expression 'morally responsible' with compatibilist or incompatibilist criteria, or if they are divided in this regard. In either case, the question would remain whether the relation between an agent and her action that grounds basic desert of blame or credit is compatible with determinism (or indeterminism). This question is substantively axiological or normative rather than conceptual, and could be raised without talk of "responsibility" given that we have a clear enough grasp of what is involved in deserving blame or credit.

Another reason for philosophers to care about what non-philosophers think is dialectical. For example, if it turned out that almost everyone had incompatibilist beliefs or intuitions—if almost everyone thought or felt that causal determination of action undermined responsibility—the compatibilist would have a more difficult time convincing people of her view; likewise for the incompatibilist if almost everyone had compatibilist beliefs. Any rationally convincing argument would need to be much more forceful than the

contrary intuitions or else be complemented with independent reasons to distrust these intuitions.[1] Moreover, to the extent that *some* epistemic weight should be given to ordinary intuitions about these issues, a position at odds with common sense would carry not only an extra dialectical burden, but also an epistemic one. As things stand, however, surveys are divided about the extent to which people are compatibilists, and even studies suggesting that one of the two positions predominates reveal a substantial minority with the opposite view, at least under some circumstances (see e.g. Nahmias et al. 2007; Nichols and Knobe 2007). Judging by mere strength of numbers, neither position is in an epistemically favorable position, and both positions face dialectical resistance.

The most direct traditional way of addressing vexed philosophical problems is to look for better arguments, to try strengthening existing ones, and to reveal problems with opposing arguments. In this context, empirical studies could help ensure that the seeming plausibility of the premises involved are more than a reflection of partisan prejudice. But as we shall see, such studies have a further potential role, in relation to so-called "error theories" offered by philosophers in acknowledgment of intuitions contrary to their own views. Such error theories include incompatibilists' suggestions that we resist incompatibilist conclusions because we do not understand how our actions are caused (Spinoza 1667/1985, 440) or because we are strongly disposed to blame-involving emotions like indignation and guilt (e.g., Nichols and Knobe 2007). On the compatibilist side, error theories include the suggestion that incompatibilist intuitions stem from a confusion of determinism with fatalism, or a confusion of causation preventing one from doing what one wants with causation generally (Hume 1739/1978, Book 2, Part 3, Section 2), or a confusion of "guidance control", which requires that one causes one's actions in a certain way, with "regulative control", which requires that one could act otherwise (Fischer 2013; cf. Fischer and Ravizza 1998). One might find it unlikely that philosophers familiar with these distinctions would in fact make those mistakes, but there is at least some reason to worry that errors afflicting folk intuitions also affect worked-out philosophical positions. Such positions are often attempts to articulate and provide justification for intuitive pre-theoretical commitments, and if these commitments are based on errors, the philosopher's view might reflect rationalizations of these pre-theoretical errors. Philosophers have rarely meant their error-theoretic proposals to do more than indicate how opposing intuitions might be mistaken. However, if empirical studies of responsibility judgments were to show that these mistakes are actually being made, and are actually at work in explaining the intuitions, this would lend much more weight to these proposals. We will examine a potential example below.

Finally, even if empirical considerations fall short of showing that some position rests on erroneous intuitions, they might nevertheless indicate that compatibilist and incompatibilist tendencies are affected by factors of unclear epistemic status. For example, studies by Shaun Nichols and Joshua Knobe (2007) suggest that subjects are considerably more willing to attribute moral responsibility to agents in a deterministic universe when asked about responsibility for a concrete action than when asked abstractly whether agents in this universe can be responsible for their actions (cf. Nahmias et al 2007). More generally, Gunnar Björnsson and Karl Persson (2012; 2013) have argued that a variety of results from experimental studies (as well as the appeal of various philosophical

arguments) can be accounted for if we (a) understand responsibility judgments as judgments attributing an explanatory relation between the agent's motivational structure and the object of responsibility, and (b) take these explanatory judgments to be selective and sensitive to explanatory interests and perspectives in much the way that everyday explanatory judgments are. Both the abstract–concrete variation and the hypothesized dependency on explanatory perspectives raise difficult methodological questions. Are abstract judgments about the possibility of responsibility more or less trustworthy than those made about concrete cases (Nichols and Knobe 2007l, 677–81)? Are judgments made from certain everyday explanatory perspectives more or less trustworthy than judgments made from explanatory perspectives made salient by abstract deterministic scenarios (Björnsson and Persson 2012, 345–8)?

We believe that experimental philosophy is relevant to the traditional debates. At the same time, it turns out to be challenging to set up experiments and interpret data in just the right way – no less difficult than negotiating traditional philosophical arguments. Both routes are valuable, but so far neither promises a way to secure significant agreement among the competing parties. To illustrate, we focus on three sorts of issues. In the following sections, we discuss an error theory for incompatibilist intuitions proposed by Eddy Nahmias and colleagues, the role that empirical studies might have in the assessment of manipulation arguments for incompatibilism, and the suggestion that empirical studies reveal that core criteria for moral responsibility ought not to be applied invariantly across different sorts of cases.

3. An error theory for incompatibilist intuitions

Nahmias's compatibilist error theory for why many subjects provide incompatibilist answers in various surveys is that they assume that in the deterministic scenario agents have no causal role in producing their actions. In his terminology, these subjects are assuming that determinism issues in the *bypassing* of agential processes such as conscious deliberation in the production of action. It would be agreed by philosophers who participate in the debate that the mere fact that an action is causally determined by factors beyond an agent's control does not preclude her deliberation, say, from playing a causal role in bringing about her actions. Thus while the assumption that determinism involves bypassing would tend to yield non-responsibility intuitions in deterministic cases, both compatibilists and incompatibilists would agree that a non-responsibility intuition with this etiology does not count against compatibilism.

Care must be taken in formulating the bypassing hypothesis, since it turns out that various candidates are apt to suggest a claim that does not amount to bypassing (Björnsson and Pereboom 2014). For example, consider one recent formulation by Nahmias:

> In general, an agent's mental states and events—her beliefs, desires, or decisions—are bypassed when the agent's actions are caused in such a way that her mental states do not make a difference to what she ends up doing. (2011, 561)

Characterizing bypassing in terms of the failure of difference-making is subject to this sort of worry. On the one hand, difference-making can be understood in terms of nomological or causal dependence. On this reading, an agent's judgment as to which action would be best makes a difference to whether an action occurs just in case her making that judgment implies, by causal law and relevant facts about the situation, that the action will occur, while the non-occurrence of the judgment implies that the action would not result (Hume 1748/2000; Lewis 1973). If subjects believe that such difference-making is ruled out by determinism, they've misunderstood determinism. On the other hand, according to traditional incompatibilism, because propositions detailing the natural laws and the remote past entail propositions describing every subsequent event, and we can't make propositions about the laws and the remote past false, we can't make a difference as to whether any such event occurs. This is the intuition that is spelled out by the Consequence Argument (van Inwagen 1983; Ginet 1990), and it invokes a more demanding, but entirely legitimate, sense of difference-making. In this second sense, difference-making requires that the difference-maker is not itself causally determined by anything else—that it provide a kind of independent input into the unfolding universe. Call this "ultimate" difference-making. If subjects are asked whether an agent's beliefs, desires, or decisions can make a difference to whether their actions will occur given determinism, ultimate difference-making might well come to mind. If an incompatibilist response is then made, it can't justifiably be set aside on the ground that the subject erroneously assumes that determinism involves bypassing.

While Nahmias did not use the difference-making formulation in his surveys, the formulations he did employ are arguably subject to similar problems. To test the bypassing hypothesis, Nahmias and his collaborator Dylan Murray (Nahmias and Murray 2010; Murray and Nahmias 2014) had subjects read different descriptions of a deterministic universe and then rate three statements about the possibility of moral responsibility and free will in that universe on a six-point scale (strongly disagree, disagree, somewhat disagree, somewhat agree, agree, strongly agree), and five statements designed to capture whether the agents' capacities for deliberative control of actions were bypassed, again on a six-point scale. Composite scores for each group of statements (*responsibility* and *bypassing*) were calculated for each subject. Interestingly, there was a strong overall correlation between scores for *bypassing* and scores for *responsibility*. Provided that ratings of statements reliably tracked subjects' attributions of moral responsibility and their belief that deliberative control was bypassed, the bypassing hypothesis would be vindicated: incompatibilist intuitions would seem to depend on the erroneous assumption that determinism involves bypassing.

There are, however, reasons to doubt that the statements designed to track belief in bypassing actually did just that (Björnsson and Pereboom 2014). The following statements are representative of those the subjects read:

> NO CONTROL: In Universe A, a person has no control over what they do.
> DECISIONS: In Universe A, a person's decisions have no effect on what they end up being caused to do.

WANTS: In Universe A, what a person wants has no effect on what they end up being caused to do.
BELIEVES: In Universe A, what a person believes has no effect on what they end up being caused to do.

Begin with NO CONTROL. The notion of control intended by Nahmias and Murray is one aligned with the nomological-dependence notion of difference-making. But there is also a notion of control corresponding to that of ultimate difference-making. It isn't confused to think that our beliefs, desires or decisions have no such ultimate control in a deterministic system. (Philosophers concerned with free will and moral responsibility often distinguish such control from compatibilist-friendly varieties; see, for example, Fischer and Ravizza's (1998) distinction between regulative and guidance control.)

DECISIONS, WANTS and BELIEVES are open to the same dual interpretations as "difference making" and "control". On one reading, A has an effect on B insofar as B is nomologically dependent on A. On another, what is required is that A is an ultimate difference-maker for B. If subjects accept DECISIONS, WANTS and BELIEVES because they deny that human decisions, desires and beliefs are ultimate difference-makers in a deterministic universe, this does not show that they confusedly take determinism to imply bypassing.

It may be, then, that the four statements designed to test for bypassing can be plausibly understood in ways allowing that determination of actions *passes through* rather than *bypasses* agents' decisions, desires and belief. We might test whether subjects' actual interpretations are indeed *throughpass*-friendly. Two surveys by Gunnar Björnsson (2014) designed to test the robustness of Nahmias and Murray's results replicated some of them: scores for statements similar to DECISIONS, WANTS and BELIEVES were strongly negatively correlated with responsibility scores. But consider the following statement, designed to specify clearly that the agent's deliberation is not bypassed:

THROUGHPASS: In Universe A, when earlier events cause an agent's action, they do so by affecting what the agent believes and wants, which in turn causes the agent to act in a certain way.

In both surveys, subjects gave scores well over the midline to statements like THROUGHPASS. This suggests that few subjects understood determinism as implying that agents' beliefs and desires are bypassed. In addition, there was no negative correlation between THROUGHPASS and *bypassing* scores, contrary to what one would have expected if subjects had interpreted DECISIONS, WANTS, and BELIEVES as implying that determination bypasses rather than passes through the agent's deliberation. These results, which were robust across different scenarios and different formulations of THROUGHPASS, strengthen the suspicion that subjects' high scores on Nahmias and Murray's bypass statements are dependent on the kinds of throughpass-compatible interpretations sketched above (for details, see Björnsson 2014).

Additional evidence against the bypassing hypothesis comes from David Rose and Shaun Nichols (2013), who criticize it based on statistical analysis of data from studies like those

of Nahmias and Murray. Nahmias and Murray noted strong correlations between *responsibility* and *bypassing* scores, just as one would expect if variations in deterministic scenarios affected responsibility attributions by affecting bypassing judgments. But a strong correlation is compatible with a variety of hypotheses about how the variables are causally related, with three of many possible alternatives illustrated in Figure 1, with arrows indicating causal relationships between variables. The first is the possibility suggested by Nahmias and Murray: variations in deterministic scenarios cause variations in subjects' beliefs that agency is bypassed in the scenarios, and such beliefs explain why subjects are reluctant to attribute responsibility. The second possibility, "Responsibility First", takes variations in scenarios to affect attributions of responsibility, and lower attributions of responsibility cause subjects' sense that agency is bypassed. The third possibility, "Common Cause", denies both these causal relations between responsibility and bypassing judgments. Instead, it postulates some factor that is affected by variations in scenarios and itself affects responsibility and bypassing judgments in opposite ways, thus explaining their negative correlation. (We indicate a possible factor common cause below.) Rose and Nichols analyzed data from a variation on the Nahmias and Murray study and found that it fit the Responsibility First model much better than the Bypass model. Moreover, these results seem stable, as they have been independently replicated in two further studies (Björnsson 2014).

All in all then, there are strong reasons to reject the bypass hypothesis. Subjects accepting *bypassing* statements need not have misunderstood determinism, and corresponding *bypassing* judgments seem to have little influence on *responsibility* judgments. This leaves the question of why *responsibility* and *bypassing* scores are consistently negatively correlated, and a better understanding of the correlation might tell us more about worries raised by determinism. Here we suggest that there might be an independently motivated explanation of this correlation: Both throughpass-friendly interpretations of *bypassing* statements and low scores on *responsibility* are explained by subjects' salient explanatory perspectives.

To see how this would work, begin with the interpretation of bypassing statements. Here we can assume that the choice between the two available interpretations is guided by considerations that are salient for the particular subject. Moreover, we can safely assume that notions like "having an effect," "having control over," or "making a difference to" are causal or explanatory notions, expressive of subjects' take on what explains what. Given these two plausible assumptions, the relative salience of the two proposed interpretations likely depends on what explanatory perspective or explanatory model is more salient for that subject. For subjects who understand *bypassing* statements in terms of ultimate difference-making, an explanatory model in which only ultimate difference-makers figure as explanatory variables will be particularly salient. For subjects who instead understand these statements as they are intended by Nahmias and Murray, an explanatory model in which agent's decisions, desires and beliefs figure as explanatory variables will be more salient.

Turning from the interpretation of *bypassing* statements to attributions of responsibility, there are independent reasons to think that the latter too are affected by the salience of

explanatory models. Björnsson and Persson (2012; 2013) argue that the ordinary notion of moral responsibility is itself an explanatory notion, such that to take an agent to be responsible for an event (an action or outcome) is to see the event as explained in a normal way by the agent's motivational structure (roughly, the agent's quality of will, or reasons-responsive mechanisms). More specifically, Björnsson and Persson argue that subjects who take determinism to undermine moral responsibility are those for whom the explanatory perspective of ordinary folk psychology is overshadowed by a deterministic perspective in which human agency is a mere dependent variable. But given what we just said about the interpretation of bypassing statements, these are just the subjects who should be (a) more inclined to interpret *bypassing* statements as concerned with ultimate difference-making, and so (b) more inclined to agree with these statements. If this is correct, this explanatory perspective would be a common cause of low responsibility attributions and high agreement with *bypassing* statements, and thus would straightforwardly account for the negative correlation between *responsibility* and *bypass* scores.[2]

4. Manipulation

Knobe and Doris (2010) point out that one prominent strand in the contemporary debate between compatibilists and incompatibilists involves devising scenarios in which ordinary intuitions will tend to diverge from what the opponent's theory predicts. In this section, we discuss problems and prospects for using empirical studies to undermine or support strategies of this kind, focusing on the main contemporary incompatibilist instance of the strategy, the manipulation argument (Taylor 1974; Ginet 1990; Pereboom 1995, 2001, 2014; Kane 1996; Mele 2006). Such an argument begins with the intuition that if a subject is causally determined to act by other agents, for example by neuroscientists who manipulate her brain, then she is not morally responsible for that action, and this is so even if she satisfies the main compatibilist conditions on moral responsibility. It continues by arguing that there are no differences between cases like this and otherwise similar ordinary deterministic examples that can justify the claim that while an agent is not morally responsible when manipulated by other agents, she can nevertheless be morally responsible in the ordinary deterministic cases.

Mele (2006) develops an elegant manipulation argument, involving only one 'original design' manipulation case, in which a goddess Diana determines Ernie's zygote so that he will at some point commit an immoral act. The challenge for the compatibilist is to point out a relevant and principled difference between this manipulation scenario and an ordinary deterministic case that would show why the agent might be morally responsible in the ordinary case but not in the manipulation example. Advocates of this manipulation argument argue that this cannot be done.

Pereboom's multiple-case manipulation argument, which has been subjected to a number of experimental studies (e.g., Sripada 2012; Feltz 2103; Murray and Lombrozo 2015), sets out several manipulation examples, the first of which features the most radical sort of manipulation consistent with the proposed compatibilist conditions. The subsequent cases are progressively more like a final example, which the compatibilist might envision to be

ordinary and realistic, in which the action is causally determined in a natural way. A challenge for the compatibilist is to point out a relevant and principled difference between any two adjacent cases that would show why the agent might be morally responsible in the later example but not in the earlier one.

Specifically, in each of the four cases Plum decides to kill White for the sake of some personal advantage, and succeeds in doing so. The action under consideration, then, is his decision to kill White—a basic mental action. This action fits certain compatibilist conditions proposed by David Hume: it is not out of character, since for Plum it is generally true that selfish reasons weigh heavily—too heavily when considered from the moral point of view—while in addition the desire that motivates him to act is nevertheless not irresistible for him, and in this sense he is not constrained to act (Hume 1739/1978). The action also meets the compatibilist condition proposed by Harry Frankfurt (1971): Plum's effective desire (i.e., his will) to murder White conforms appropriately to his second-order desires for which effective desires he will have. That is, he wills to murder her, and he wants to will to do so. In addition, the action satisfies the reasons-responsiveness condition advocated by John Fischer and Mark Ravizza (1998): Plum's desires can be modified by, and some of them arise from, rational consideration of his reasons, and if he believed that the bad consequences for himself that would result from his killing White would be more severe than he actually expects them to be, he would not have decided to kill her. This action also satisfies the related condition advanced by Jay Wallace (1994): Plum has the general ability to grasp, apply, and regulate his actions by moral reasons. For instance, when egoistic reasons that count against acting morally are weak, he will typically act for moral reasons instead. This general ability provides him with the capacity reflectively to revise and develop his moral character and commitment over time, and for his actions to be governed by those moral commitments, a condition that Mele (1995; 2006) and Haji (1998; 2009) underscore. Supposing that Plum is causally determined by factors beyond his control to decide as he does, is it plausible that he is morally responsible for his decision?

The four cases exhibit varying ways in which Plum's decision to kill White might be causally determined by factors beyond his control. In Case 1 (Pereboom 2014, 76-77 version), a team of neuroscientists has the ability to manipulate Plum's neural states at any time by radio-like technology. On this particular occasion, they do so by pressing a button just before he begins to reason about his situation, which they know will produce in him a neural state that realizes a strongly egoistic reasoning process, which the neuroscientists know will deterministically result in his decision to kill White (cf. Shabo 2010). Plum would not have killed White had the neuroscientists not intervened, since his reasoning would then not have been sufficiently egoistic to produce this decision. His reasoning is consistent with his character because it is frequently egoistic and sometimes strongly so. Still, it is not in general exclusively egoistic, because he sometimes successfully regulates his behavior by moral reasons, especially when the egoistic reasons are relatively weak.

In Case 2, Plum is just like an ordinary human being, except that a team of neuroscientists programmed him at the beginning of his life so that his reasoning is often but not always egoistic (as in Case 1), and at times strongly so, with the intended consequence that in his current circumstances he is causally determined to engage in the egoistic reasons-

responsive process of deliberation and to have the set of first and second-order desires that result in his decision to kill White. In Case 3, Plum is an ordinary human being, except that the training practices of his community causally determine the nature of his deliberative reasoning processes so that they are frequently but not exclusively rationally egoistic (the resulting nature of his deliberative reasoning processes are exactly as they are in Cases 1 and 2). This training was completed before he developed the ability to prevent or alter these practices. Finally, in Case 4, everything that happens in our universe is causally determined by virtue of its past states together with the laws of nature. The neural realization of Plum's reasoning process and decision is exactly as it is in Cases 1-3; he has the general ability to grasp, apply, and regulate his actions by moral reasons, and it is not because of an irresistible desire that he decides to kill.

Pereboom claims that there are no differences between adjacent cases that would justify the claim that Plum is not responsible in the earlier but not in the later case. In each, Plum satisfies the prominent compatibilist conditions on moral responsibility. In each the neural realization of his reasoning process and decision is the same, although the causal histories of these realizations differ.

One widespread compatibilist hypothesis is that a distinguishing feature of the ordinary deterministic case is that the causal determination of Plum's decision is not brought about by other agents (Lycan 1997). The key claim is that what is generating the non-responsibility and non-free-will intuitions in the first three cases is not causal determination per se, but causal determination by other agents. Adam Feltz (2013) as well as Dylan Murray and Tania Lombrozo (2015) have tested this suggestion. Feltz found diminished judgments of moral responsibility in cases of causal determination by other agents relative to naturalistic determination. But only if the manipulation by other agents was intentional and direct, as in Case 1, did his subjects, on average, fall below the midpoint between 'strongly agree' and 'strongly' disagree. On Murray and Lombrozo's interpretation of their results, they indeed show that intentional control by other agents robustly generates intuitions of absence of responsibility and free will, while causal determination per se does not. They conclude that because causal determation per se does not robustly generate such intuitions, the comparison to manipulation doesn't support incompatibilism.

One worry about the conclusions of Feltz, Murray and Lombrozo concerns the fact that the terms 'moral responsibility' and 'free will' are multiply ambiguous, and that according to the incompatibilist, only one central pair of notions of 'free will' and 'moral responsibility' gives rise to an incompatibility with causal determination of action by factors beyond the agent's control, while others are compatible with causal determination. Again, in the historical debate, the variety of free will at issue is the sort required for moral responsibility in a particular but pervasive sense, set apart by the notion of basic desert. Rejecting this kind of moral responsibility leaves other senses intact, for example, a forward-looking answerability notion that aims at protection, reconciliation, and moral formation. Our actual practice features this forward-looking sense, and likely others as well. When we ask experimental subjects whether an agent described in some scenario is morally responsible, all of these senses are potentially in play. According to the incompatibilist, if the manipulation examples are set up appropriately, then the intuition in

all these cases should be that the agent is not morally responsible in the basic desert sense, but is morally responsible in the forward-looking sense just set out. Perhaps most crucially, the agent, by virtue of being reasons-responsive, will be disposed to moral improvement upon being blamed. But then if in our surveys we don't distinguish such senses of responsibility, the incompatibilist hypothesis isn't being adequately tested. In fact, if asked whether Plum in Case 1 is morally responsible without factoring out the different senses, even the incompatibilist author of this contribution would respond that he is (Pereboom 2014, 136). As a corrective, experimental prompts might differentiate among different senses of moral responsibility. Note that while Murray and Lombrozo's study also asked subjects whether the agent in question deserved blame, even this is ambiguous between crucially distinct notions: basic desert and desert derived from consequentialist or contractualist considerations. Incompatibilists might be disposed to agree that manipulated or causally determined agents can deserve blame in the derived sense.

Feltz found that subjects tend to agree more strongly with judgments of moral responsibility and free will for Plum as we move from Case 1 to Case 4, and Murray and Lombrozo's study yielded similar results. However, these findings do not contravene the incompatibilist's expectations. The strategy of the manipulation argument does not involve claiming that subjects' immediate assessments of freedom and moral responsibility will be the same in the four cases, but rather that there is no difference among the cases that can explain such variations in moral responsibility assessments in a principled way. In fact, the incompatibilist predicts that immediate assessments about responsibility will generally differ between Cases 1 and 4, but maintains that at this point a further phase of the argument becomes pertinent: a request to explain the differences in intuition in a principled way. *Should* intentional direct manipulation by other agents make a difference relative to natural causal determination in assessing basic desert moral responsibility? It would be valuable to survey respondents taking these considerations into account. Subjects might be challenged to explain differential judgments across the cases, and then tested to see whether their judgments about the individual cases change as a result.

The incompatibilist might also object that despite the reactions of the subjects, the difference between manipulation by another agent and naturalistic determination might still be irrelevant to moral responsibility in the basic desert sense. One might test this hypothesis by having subjects imagine further cases that are exactly the same as Case 1 or Case 2, except that states at issue are instead produced by a spontaneously generated machine—a machine specified to have no intelligent designer (Pereboom 2001) or a force field (Mele 2006). However, it's hard to separate the idea of a sophisticated machine from intelligent, intentional, designers of that machine, even if it's specified that the machine is spontaneously generated. The mechanism by which a force field manipulates may be too unclear, and it might well suggest bypassing to at least some subjects.

In response to such concerns, Björnsson (2015) constructed a scenario where all standard compatibilist conditions on responsibility are satisfied but where a non-agential cause—an infection—slowly turns the agent increasingly egoistic without bypassing or undermining his agential capacities. Based on the hypothesis that subjects take responsibility to be undermined when they understand the agent from an explanatory perspective in which the

agent's deliberation is a mere dependent variable (Björnsson and Persson 2012; 2013), he predicted that if subjects were prompted to see the agent's behavior as dependent on this non-agential cause, this would undermine attributions of responsibility to roughly the same extent as the introduction of an intentional manipulator. This was indeed the case: in a study involving 416 subjects, the infection undermined attributions of free will and moral responsibility to the same degree as indoctrination cases of intentional manipulation. This study suggests that incompatibilists might be able to employ the same generalization strategy used in manipulation arguments without introducing intentional external control, and thus without being subject to the experimentally-driven objections developed by Feltz, Murray and Lombrozo.[3]

5. Variantism and invariantism

A number of philosophers argue that the results of surveys provide confirming evidence for a meta-view about moral theories, *variantism*. The dominant countervailing position claims that the core criteria for moral responsibility ought to be applied invariantly across all cases. Variantism holds that this is not so, and that there are substantial respects in which core criteria ought to be applied differently depending on the circumstances (Knobe and Doris 2010).

One relevant hypothesis tested by Nichols and Knobe (2007) is that subjects tend toward incompatibilism when the scenario described is abstract and general, but toward compatibilism when it is concrete and vivid. Subjects were presented with an account of a universe – Universe A -- in which all events unfold in accord with deterministic laws. The abstract question was:

> In Universe A, is it possible for a person to be fully morally responsible for their action?

The concrete question was:

> In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and three children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down his house and kills his family. Is Bill fully morally responsible for killing his wife and children?

In the abstract condition, 14% of the subjects agreed that it is possible for a person to be fully morally responsible for their action in the specified circumstances, while in the concrete condition 72% of the subjects affirmed that Bill is fully morally responsible for what he did.

Nichols and Knobe canvas several possible explanations for this variation. One could, for instance, attribute the high-affect response to the distorting effect of emotion. But one

might instead think it to be suggestive of variantism, whereupon the concept of moral responsibility ought to be applied differently under varying conditions of affect. Knobe and Doris (2010) address the objection that it's just obvious that the high-affect/low-affect survey shows us nothing about how we ought to apply moral responsibility concepts. Their response is:

> The fact that a particular view strikes people as obvious does not show us anything about the nature of the competence underlying ordinary attributions of moral responsibility. What would show us something about the nature of competence is a specific, testable model that accounts for the existing data and can then be used to generate new predictions that can be examined in further studies. (2010, 348)

If this is right, then there is a massively important role for experimental philosophy in determining how we ought to apply our responsibility concepts.

One line of defense against variantism is developed by Dana Nelkin (2007). She argues that the degree of variation that studies reveal can often be accounted for by invariantist accounts. Consider the abstract/concrete variation and whether it can be accounted for by an invariantist theory in which moral responsibility is aligned with the ability to effectively deliberate in accord with the relevant reasons. Nelkin proposes that a significant proportion of the population may at least initially assume that determinism rules out the possibility of actions resulting from such a process, and instead consigns causation of action to mechanical factors such as neural states (Nahmias 2006). She suggests that in the concrete case this assumption may tend to be overridden by the description of the way the action came about. Indeed, Nichols' and Knobe's vignette involving Bill includes a vivid description of the deliberative reasoning process that results in his decision to kill his wife and children. In the abstract case, the vignette did not include a description of a reasoning process, and this might explain why the assumption is not overridden. More generally, Nelkin's strategy counsels that we derive apparently varying judgments from an invariantist theory together with natural but perhaps unjustified theoretical and empirical assumptions.[4]

Nelkin also proposes that sometimes apparently varying judgments can be derived from the invariantist theories themselves without any controversial empirical judgments. Consider a survey that confirms an apparent variance in how subjects judge those who act with a great deal of emotion. David Pizarro et al. (2003) presented one group of subjects with a vignette about a morally good action: "Because of his overwhelming and uncontrollable sympathy, Jack impulsively gave the homeless man his only jacket even though it was freezing outside," and another group with this vignette about a bad action: "Because of his overwhelming and uncontrollable anger, Jack impulsively smashed the window of the car parked in front of him because it was parked too close to his," Contrasting cases were also presented in which the agent acted 'calmly and deliberately." Subjects judged agents much less blameworthy when they acted badly with emotion relative to acting badly without. But in the case of good action the difference was negligible. Nelkin suggests, however, that this difference is explained by an invariantist theory according to which moral responsibility aligns with the ability to act in accord with good

reasons (Wolf 1990; Nelkin 2007; 2011). In the good case, the emotion tends to highlight the good reason, while in the bad case the emotion obscures it, and thus can be seen as an excuse.

A second line of resistance involves advancing the claim that some fundamental invariantism is a feature of the ground rules of morality, and is significantly independent of empirical testing. Consider the well-known study of sentencing practices of Israeli judges, in which Danziger, Levav, and Avnaim-Passo (2011) surveyed rulings judges made during three subsequent daily decision sessions, each of which was followed by food breaks. They found that "the percentage of favorable rulings drops gradually from ≈65% to nearly zero within each decision session and returns abruptly to ≈65% after a break." Here it might strike one as obvious that such a pattern does not reflect competence. It seems clear that further empirical testing is not required to determine whether it is. This verdict might of course be reflected by apriorist Kantian theory, but even Hume, who grounds morality in sympathy and sentiment, allows for such an a priori element. Sympathy and sentiment is variable: "nor can I feel the same lively pleasure from the virtues of a person, who liv'd in Greece two thousand years ago, that I feel from the virtues of a familiar friend and acquaintance. Yet I do not say that I esteem the one more than the other" (1739/1978, 581). Hume's solution is that when we make moral judgments "we fix on some steady and general points of view; and always, in our thoughts, place ourselves in them, whatever may be our present situation" (1739, 582). On this conception, there is a degree of invariantism built into the ground rules for moral judgment and, more specifically, for the attribution of moral responsibility. Suppose that studies found systematic racial bias in sentencing. Very plausibly, no studies could show that such racial bias reflects competence, and no studies are needed to show that it reflects incompetence. On one diagnosis, we know this by understanding the ground rules of morality. But even if this is so, it's open that experimental surveys are valuable insofar as they can help to determine that there is some degree of variation in how we ought to attribute moral responsibility, and where that variation exists.

6. Final words

We conclude that it's currently unclear what upshot empirical surveys have for the assessment of the bypassing error theory for incompatibilist intuitions, for defeating manipulation arguments for incompatibilism, and for confirming variantism about responsibility criteria. In each of these cases, there are significant problems for setting up effective surveys and for interpreting data in convincing ways. These difficulties seem no less challenging than in the case of traditional philosophical arguments. We propose that both routes to philosophical clarification are nonetheless valuable, even though neither has yet been able to secure significant agreement among opposing camps.[5]

[1] Nahmias et al. (2006, 30–32) argue that incompatibilism is in particular need of intuitive support given that it postulates metaphysically stronger requirements on responsibility. But one might also think that what is in particular need of justification are claims that some people deserve to be treated better or worse than others. This would put a greater burden of justification on compatibilism, as it postulates weaker restrictions on when blame and credit are deserved.

[2] For a development of this explanation, see Björnsson 2014. Rose and Nichols (2013) propose an alternative explanation of the negative correlation between *responsibility* and *bypass* scores. Their suggestion, further pursued in Chan, Deutsch and Nichols (2015), is that subjects (a) take free will to be necessary for the existence of beliefs, desires and decisions, and (b) accept bypassing statements when they take determinism to rule out free will and thus rule out the existence of such states: if there are no decisions, decisions have no effect on what agents do. For criticism and experimental evidence against this interpretation, see Björnsson 2014.

[3] Following Björnsson and Persson (2012, 345–48), Björnsson (2015) instead suggests that there might be general methodological reasons not to rely on these intuitions.

[4] Björnsson and Persson (2013) can be seen as generalizing this strategy to a wider range of phenomena.

[5] Björnsson's work on this chapter was supported by a grant from the John Templeton Foundation as well as one from Riksbankens Jubileumsfond. Their views are not necessarily reflected by the opinions expressed in this chapter.

Bibliography

Ayer, Alfred J. 1954. "Freedom and Necessity." In *Philosophical Essays*, edited by Alfred J. Ayer, 271–84. London: Macmillan.

Björnsson, Gunnar. 2014. "Incompatibilism and 'Bypassed' Agency." In *Surrounding Free Will*, edited by Alfred Mele, 95–122. New York: Oxford University Press.

Björnsson, Gunnar. 2015. "Manipulators, Parasites, and Generalization Arguments." In preparation.

Björnsson, Gunnar, and Karl Persson. 2012. "The Explanatory Component of Responsibility." *Noûs* 46 2: 326–54. DOI: 10.1111/j.1468-0068.2010.00813.x

Björnsson, Gunnar and Karl Persson. 2013. "A Unified Empirical Account of Responsibility Judgments." *Philosophy and Phenomenological Research* 87 3: 611–39. DOI: 10.1111/j.1933-1592.2012.00603.x

Björnsson, Gunnar, and Derk Pereboom. 2014. "Free Will Skepticism and Bypassing." In *Moral Psychology, Vol. 4*., edited by Walter Sinnott-Armstrong, 27–35. Cambridge, MA: MIT Press.

Chan, Hoi-yee, Max Deutsch, and Shaun Nichols. 2015. "Free Will and Experimental Philosophy." This volume.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences* 108 17: 6889–92. DOI: 10.1073/pnas.1018033108.

Feltz, Adam. 2013. "Pereboom and Premises: Asking the Right Questions in the Experimental Philosophy of Free Will." *Consciousness and Cognition* 22 1: 53–63. DOI: doi:10.1016/j.concog.2012.11.007

Fischer, John, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, John. 2013. "The Frankfurt Style Cases: Philosophical Lightning Rods." In *Free Will and Moral Responsibility*, edited by Ish Haji and Justin Caouette, 43–57. Newcastle: Cambridge Scholars Publishing.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 1: 5–20.

Ginet, Carl. 1990. *On Action*. Cambridge: Cambridge University Press.

Haji, Ishtiyaque. 1998. *Moral Appraisability,* New York: Oxford University Press.

Haji, Ishtiyaque. 2009. *Incompatibilism's Allure: Principal Arguments for Incompatibilism*. Peterborough ON: Broadview Press.

Hobart, R. E. 1934. "Free Will as Involving Determinism and Inconceivable without It." *Mind* 43 169: 1–27.

Hume, David. 1739/1978. *A Treatise of Human Nature*. Oxford: Oxford University Press.

Hume, David. 1748/2000. *An Enquiry Concerning Human Understanding*, Oxford: Oxford University Press.

Kane, Robert. 1996. *The Significance of Free Will,* New York: Oxford University Press.

Knobe, Joshua. 2003. "Intentional Action in Folk Psychology: An Experimental Investigation." *Philosophical Psychology* 16 2: 309–24.

Knobe, Joshua, and John Doris. 2010. "Responsibility." In *The Moral Psychology Handbook*, edited by John Doris and the Moral Psychology Research Group, 321–54. Oxford: Oxford University Press.

Lewis, David. 1973. "Causation." *The Journal of Philosophy* 70 17: 556–67.

Lycan, William. G. 1997. *Consciousness*. Cambridge, MA: MIT Press University Press.

McKenna, Michael. 2012. *Conversation and Responsibility*, New York: Oxford University Press.

Mele, Alfred. 1995. *Autonomous Agents: From Self-Control to Autonomy*, Oxford: Oxford University Press.

Mele, Alfred. 2006. *Free Will and Luck*, New York: Oxford University Press.

Mele, Alfred. 2009. *Effective Intentions,* New York: Oxford University Press.

Murray, Dylan and Eddy Nahmias. 2014. "Explaining Away Incompatibilist Intuitions." *Philosophy and Phenomenological Research* 88 2: 434–67.

Murray, Dylan and Tania Lombrozo (2015). "Effects of Manipulation on Attributions of Causation, Free Will, and Moral Responsibility." Under review.

Nahmias, Eddy. 2006. "Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism." *Journal of Cognition and Culture* 6 1–2: 215–37.

Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. 2006. "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research* 73 1: 28–53.

Nahmias, Eddy, Justin D. Coates, and Trevor Kvaran. 2007. "Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions." *Midwest Studies In Philosophy* 31 1: 214–42.

Nahmias, Eddy, and Dylan Murray. 2010. "Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions." In *New Waves in Philosophy of Action*, edited by Jesús Aguilar, Andrei Buckareff, and Keith Frankish, 112–29. New York: Palgrave-Macmillan.

Nahmias, Eddy. 2011. "Intuitions about Free Will, Determinism, and Bypassing." In *The Oxford Handbook of Free Will*, 2nd ed., edited by Robert Kane, 555–76. New York: Oxford University Press.

Nelkin, Dana. 2007. "Do We Have a Coherent Set of Intuitions About Moral Responsibility?" *Midwest Studies in Philosophy* 31 1: 243–59.

Nelkin, Dana. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Nichols, Shaun, and Joshua Knobe. 2007. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Noûs* 41 4: 663–685.

Pereboom, Derk. 1995. "Determinism *Al Dente*." *Noûs* 29 1: 21–45.

Pereboom, Derk. 2001. *Living without Free Will*. Cambridge: Cambridge University Press.

Pereboom, Derk. 2013. "Free Will Skepticism, Blame, and Obligation." In *Blame: Its Nature and Norms*, edited by Neal Tognazzini and D. Justin Coates, 189–206. New York: Oxford University Press.

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

Pizarro, David, Eric Uhlmann, and Peter Salovy. 2003. "Asymmetry in Judgments of Moral Blame and Praise: The Role of Perceived Metadesires." Psychological Science 14 3: 267–72.

Rose, David and Shaun Nichols. 2013. "The Lesson of Bypassing." *Review of Philosophy and Psychology* 4 4: 599–619.

Sartorio, Carolina. 2014. "The Problem of Determinism and Free Will Is Not the Problem of Determinism and Free Will." In *Surrounding Free Will*, edited by Alfred Mele, 255–73. New York: Oxford University Press.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Shabo, Seth. 2010. "Uncompromising Source Incompatibilism." *Philosophy and Phenomenological* Research 80 2: 349–83.

Smith, Angela. 2008. "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138 3: 367–92.

Spinoza, Baruch. 1677/1985. *Ethics*. In *The Collected Works of Spinoza*, Vol. 1, edited and translated by Edwin Curley. 401-617. Princeton NJ: Princeton University Press.

Sripada, Chandra. 2012. "What Makes a Manipulated Agent Unfree?" *Philosophy and Phenomenological Research* 85 3: 563–93.

Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Oxford University Press.

Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75 1: 5–24.

Strawson, Peter F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48: 187–211.

Taylor, Richard. 1974. *Metaphysics,* 4th ed. Englewood Cliffs: Prentice-Hall.

van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

Vargas, Manuel. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Wolf, Susan. 1990. *Freedom within Reason*. Oxford: Oxford University Press.

List of figures, with caption:

Figure 1: Three causal models