Free Will Skepticism and Bypassing

Moral Psychology, v. 4, W. Sinnott-Armstrong, ed., Cambridge: MIT Press, 2014.

Penultimate draft

Gunnar Björnsson and Derk Pereboom

Two routes to the claim that free will is an illusion—free will skepticism—feature prominently in the current discussion. A first, which denies the causal efficacy of the types of willing required for free will, receives its contemporary impetus from certain kinds of studies in neuroscience, pioneered by Benjamin Libet and Daniel Wegner. A second, found especially in the philosophical literature, does not deny the causal efficacy of the will but instead claims that whether this causal efficacy is deterministic or indeterministic, it does not achieve the level of control to count as free will by the standards of the historical debate. 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*. For an agent to be morally responsible for an action in this sense is for it to be hers in such a way that she would deserve to be the recipient of an expression of moral indignation if she understood that it was morally wrong, and she would deserve to be the recipient of an expression of praise if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve to be the recipient of the expression of such an attitude just because she has performed the action, given sensitivity to its moral status,

and not, for example, merely by virtue of consequentialist or contractualist considerations (Pereboom 2001, 2012).

Rejecting this kind of moral responsibility 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. 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 2012). The main thread of the historical free will debate does not pose determinism as a challenge to moral responsibility conceived in this way, and free will skeptics can accept that we are morally responsible in this sense. Nahmias claims that most contemporary philosophers are compatibilists, and a recent survey by David Bourget and David Chalmers (2009) supports this assessment. However, some philosophers self-identify as compatibilists because they hold that determinism is compatible with our being morally responsible in some non-basic-desert sense (perhaps Frank Jackson, 1998, pp. 44–45, is an example). If this counts as compatibilism, however, virtually everyone is a compatibilist. To track the main divisions within the philosophical debate, we should not count this as compatibilism.

Nahmias sides with genuine compatibilism, according to which agents can have the sort of free will required for moral responsibility in the sense at issue even if their actions are determined by factors beyond their control. One should note that while the historical philosophical debate tends to focus on whether free will in this sense is compatible with determinism generally construed, as Nahmias notes, the more pertinent issue is whether free will is compatible with our actions being determined by factors beyond our control (cf. Sartorio, ms.). One prominent way in which this sort of compatibilism is challenged is by manipulation examples (e.g., Pereboom, 1995, 2001; Kane, 1996; Mele, 2006). This strategy begins by arguing that if a subject is causally determined to act by other agents—for example, by neuroscientists who manipulate her brain—then she is intuitively 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, can nevertheless be responsible in the ordinary deterministic examples.

The most common way to argue against the compatibility of the sort of free will at issue with *in*determinism is by a luck objection. Here is one version. Consider a decision made in a context in which moral reasons favor one action, prudential reasons favor a distinct and incompatible action, and the net strength of these sets of reasons are in close competition. On an event-causal libertarian picture, the agent-involving causal conditions antecedent to the decision would leave it open whether the decision will occur, and the agent has no further causal role in determining whether it does. With

the causal role of the antecedent events already given, whether the decision ensues is not settled by any causal factor involving the agent. In fact, given the causal role of all causally relevant antecedent events, *nothing* settles whether the decision occurs. Thus on the event-causal libertarian picture agents lack the control required for moral responsibility (Pereboom, 2001).

Nahmias's paper focuses on the first type of argument, the one inspired by the neuroscientific studies, but he also weighs in on the distinctively philosophical challenge. He first makes the distinction between naturalist views that claim that causation occurs only at the most basic level, and those that endorse higher-level causation. This issue is still hotly contested, but we agree with Nahmias that higher-level causation is defensible. Then it's what Nahmias calls modular epiphenomenalism, according to which conscious processes can in principle cause actions, but they "occur too late, or in the wrong place, to cause our actions," that poses the real threat to free will.

We largely endorse the objections Nahmias raises against the extant versions of this kind of skeptical strategy, and we will highlight several of them. One especially serious counterconsideration, invoked by Nahmias and developed in meticulous detail by Mele (2009), stems from the fact that there is no direct way to tell which conscious phenomena, if any, correspond to which neural events. In particular, in the Libet studies, it is difficult to determine what the readiness potential corresponds to—for example, is it an intention formation or decision, or is it merely an urge of some sort? If it is just an urge, and the readiness potential does not correspond to the formation of an

intention or decision, then it remains open that the intention formation or decision is a conscious event.

Moreover, almost everyone on the contemporary scene who believes we have free will, whether compatibilist or libertarian, also maintains that freely willed actions are caused by virtue of a chain of events that stretches backward in time indefinitely. At some point in time these events will be such that the agent is not conscious of them. Thus, all free actions are caused, at some point in time, by unconscious events. However, as Nahmias correctly points out, the concern for free will raised by Libet's work is that all of the relevant causing of action is (typically) nonconscious, and consciousness is not causally efficacious in producing action. Given determinist compatibilism, however, it's not possible to establish this conclusion by showing that nonconscious events that precede conscious choice causally determine action since such compatibilists hold that every case of action will feature such events, and that this is compatible with free will. And given most incompatibilist libertarianisms, it's also impossible to establish this conclusion by showing that there are nonconscious events that render actions more probable than not by a factor of 10% above chance (Soon et al., 2008) since almost all such libertarians hold that free will is compatible with such indeterministic causation by unconscious events at some point in the causal chain (De Caro, 2011).

Furthermore, Nahmias correctly notes the unusual nature of the Libet-style experimental situation, that is, one in which a conscious intention to flex at some time in the near future is already in place, and what is tested for is the specific implementation of this general decision. As he convincingly points out, it's often the

case—when, for instance, we drive or play sports or cook meals—that we form a conscious intention to perform an action of general sort, and subsequent specific implementations are not preceded by more specific conscious intentions. But in such cases the general conscious intention is very plausibly playing a key causal role. In Libet's situations, when the instructions are given, subjects form conscious intentions to flex at some time or other, and if it turns out that the specific implementations of these general intentions are not in fact preceded by specific conscious intentions, this would be just like the kinds of driving and cooking cases Nahmias cites. It seems that these objections cast serious doubt on the potential for the neuroscientific studies to undermine the claim that we have the sort of free will at issue in the historical debate.

Bypassing

The cornerstone in Nahmias's bulwark against incompatibilism and the skeptical threat it poses is the hypothesis that incompatibilist intuitions illegitimately presuppose that determinism involves "bypassing," that is, roughly, that determinism involves the claim that agents have no causal role in producing their actions. Given the central role this bypassing hypothesis has in Nahmias's compatibilist strategy, we will focus on it in some detail. It would be agreed by participants in the debate generally 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 nonresponsibility intuitions in deterministic cases, both compatibilists and

incompatibilists would agree that a nonresponsibility intuition with this etiology does not count against compatibilism.

However, great care must be taken in formulating the bypassing hypothesis since it turns out that various candidates express or at least are apt to suggest a claim that does not amount to bypassing. For example, consider one recent formulation by Nahmias (2011):

> 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. (p. 561)

Characterizing bypassing in terms of the failure of difference making is subject to this sort of concern. 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, say, makes a difference to whether an action occurs just in case the agent's making that judgment implies, by causal law and relevant facts about the situation, that the action will occur, whereas the nonoccurrence of the judgment implies that the action would not occur (Hume, 1748; Lewis, 1973). If people think that such difference making is ruled out by determinism, they've misunderstood determinism. On the other hand, traditional incompatibilism has it that because propositions detailing the natural laws and the remote past entail propositions describing every subsequent event, and agents can't render propositions about the laws and the remote past false, agents cannot make a difference to whether any such event occurs. This is the intuition

that is spelled out by the Consequence Argument (van Inwagen 1983), and it invokes a more demanding, but perfectly legitimate, sense of difference making. In this sense, difference making requires that the difference maker is *an independent variable* in the causal system of the universe, that is, a variable the value of which is not determined by the value of other variables in that system. Call this "ultimate" difference making. If subjects are asked whether an agent's beliefs, desires, or decisions can make a difference whether their actions occur given determinism, this second sense might come to mind—especially among subjects who take the absence of such difference making to undermine free will. If an incompatibilist response is then generated, it can't justifiably be set aside on the ground that the subject mistakenly assumes that determinism involves bypassing.

While Nahmias did not employ the difference-making formulation in his experimental surveys, the formulations he did use are subject to similar problems. To test the bypassing hypothesis, Nahmias and his collaborator Dylan Murray (2010) had subjects read descriptions of a deterministic universe, 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 rate five statements meant 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 (*free will* and *bypassing*) were calculated for each subject. Interestingly, the overall correlation between scores for bypassing and scores for free will was very strong. 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. 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.

PAST DIFFERENT: In Universe A, everything that happens has to happen, even if what happened in the past had been different.

Start with NO CONTROL. The notion of "having control over" intended by Nahmias and Murray is presumably one corresponding to the nomological-dependence notion of difference making, a notion on which the strings can perhaps be said to have control over the marionette. However, there is also a notion of control corresponding to that of ultimate difference making: On this notion, the strings have no control over the marionette because their movement is completely dependent on the manipulator. It is not 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 sorts; see, e.g., Fischer & Ravizza's 1998 distinction between regulative and guidance control.)

DECISIONS, WANTS, and BELIEVES are open to roughly the same pair of 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, however, 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, they need not be confused about the nature of determinism.

Finally, PAST DIFFERENT also naturally allows for an interpretation that does not imply bypassing. Though we find the statement somewhat difficult to parse, we take the intended reading to be as follows:

UNIVERSAL BYPASS: For each actual event in Universe A, that event would have taken place even if prior events had been different.

Having in mind the necessitation of the deterministic scenario, however, one might well read the modal "has to happen" in PAST DIFFERENT as expressing a causal or nomological necessity, meaning roughly "follows from the past and causal laws." PAST DIFFERENT would then be understood as follows:

COUNTERFACTUALLY ROBUST DETERMINISM: Even if its past had been different, each event in Universe A would still have followed from the past and causal laws.

This clearly does not imply bypassing.

It seems to us, then, that the five statements designed to test for bypass can be plausibly understood in ways allowing that determination of actions *passes through* rather than *bypasses* agents' decisions, desires, and belief. Why think, though, that subjects' actual interpretations are "throughpass"-friendly in this way? A survey designed to test the robustness of Nahmias and Murray's results replicated some of them: Scores for statements quite similar to DECISIONS, WANTS, and BELIEVES were very strongly negatively correlated with free will scores. However, consider the following statement, designed to straightforwardly state 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.

Two groups of, altogether, 69 subjects completing the survey gave high scores overall to this and a similar statement (M = 4.17), with only 3 "strongly disagreeing" and 7 "disagreeing." This suggests that few subjects understood determinism as implying that agents' beliefs and desires are bypassed. Moreover, THROUGHPASS scores showed no meaningful correlation with free will scores (r = 0.12), suggesting that incompatibilist intuitions do not stem from mistaken bypass interpretations of determinism. Although further studies are needed to replicate and better understand these results, they strengthen the suspicion that subjects scoring high on Nahmias and Murray's bypass statements depend on the sort of throughpass-friendly interpretations sketched above. (For a discussion of such further studies, see Björnsson 2013.)

There are also more general reasons to anticipate a significant correlation between throughpass-friendly interpretations and low scores on free will. First, we should expect the choice between available interpretations to be guided by considerations salient for the particular subject. Subjects that take lack of ultimate difference making to undermine free will are more likely than others to find relevant interpretations of NO CONTROL, DECISIONS, WANTS, and BELIEVES involving such difference making. Similarly, subjects who take the necessitation of later events by earlier events to undermine free will are more likely to interpret "has to" in PAST DIFFERENT as expressing just that sort of necessity. Second, notions like "having an effect," "having control over," or "making a difference to" are explanatory notions. According to a recent account by Björnsson and Persson (2012a, 2012b), the ordinary notion of moral responsibility is itself an explanatory notion: People take an agent to be morally responsible for an object only if a relevant motivational structure of the agent is taken to be part of a significant or salient explanation of that object. Björnsson and Persson (2012b) suggest that subjects who take determinism to undermine moral responsibility are those for whom the explanatory perspective of ordinary folk psychology is trumped by a deterministic perspective in which human agency is a mere dependent variable. However, this is exactly the sort of explanatory perspective from which it makes sense to deny that humans have relevant control over their actions, or that their deliberation makes a difference or has an effect: All the relevant control, differences, and effects have their locus at the initial state of the universe. On neither of these explanations of the negative correlation between free will and bypassing scores do subjects with

incompatibilist intuitions take determinism to imply that actions fail to depend nomologically on beliefs, desires, and decisions.

Free Will and Science

Nahmias's project involves developing a naturalistic defense of free will, that is, a defense that does not stray beyond the bounds of natural science. Some of the neuroscientists he cites appear to suppose that if any sort of naturalism is true, or if all of our actions are governed by natural law, we won't have free will in the sense at issue. He correctly points out that this supposition can't simply be assumed, or thought to be a consequence of the definition of free will. However, it still may be true, given the soundness of the skeptical arguments canvassed earlier. For if the manipulation argument establishes that we don't have free will if our actions are governed by deterministic laws, and the disappearing-agent objection shows that we don't have free will if our actions are solely event caused and governed by probabilistic laws, a naturalistic account of free will (in the sense at issue in the debate) may well be ruled out.

Nahmias boldly claims that science can explain how we have free will. However, this would be true only given his controversial compatibilist assumptions. Contemporary compatibilists typically specify naturalistic and causal conditions on free will—Fischer and Ravizza (1998), for example, propose that free actions are caused by reasons-responsive processes. Natural science might well be able to explain how actions can be caused in this way. But it's controversial that this amounts to explaining how actions can be freely willed, supposing that freely willed actions are those for which agents have the

control required to be responsible in the basic desert sense. For, as noted above, it's controversial whether any naturalistic account will explain how agents can have this kind of control.

Science, all by itself, has the potential for explaining how we might be morally responsible in a forward-looking sense, one that, for example, aims the moral formation of the agents involved. It's uncontroversial that moral formation and the kind of control in action and over character it requires are causal notions, which natural science thus might well illuminate. The naturalistic credentials of basic desert are not so straightforward. The widespread *belief* that we are morally responsible in this sense might well be explained by naturalistic psychology and sociology, but a naturalistic account of our actually being responsible in this sense is a more daunting prospect.

References

<edb>Björnsson, G. (2013). Incompatibilism and 'bypassed' agency. Forthcoming in A. Mele (Ed.) Surrounding Free Will. New York: Oxford University Press.</edb>

<jrn>Björnsson, G., & Persson, K. (2012a). The explanatory component of responsibility. *Noûs*, *46*, 326–354.</jrn>

<jrn>Björnsson, G., & Persson, K. (2012b). A unified empirical account of responsibility judgments. *Philosophy and Phenomenological Research*. Epub ahead of print. doi:10.1111/j.1933-1592.2012.00603.x.</jrn>

<eref>Bourget, D., & Chalmers, D. (2009). PhilPapers Survey, http://philpapers.org/surveys/</eref>

<edb>DeCaro, M. (2011). Is emergentism refuted by the neurosciences? The case of free will. In A. Corradini & T. O'Connor (Eds.), *Emergence in Science and Philosophy* (pp. 190–21). London: Routledge.</edb>

<bok>Fischer, J., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.</bok>

<bok>Hume, David. (1748). *An enquiry concerning human understanding*.</bok>

<bok>Jackson, F. (1998). *From metaphysics to ethics*. Oxford: Oxford University Press.</bok>

<bok>Kane, R. (1996). *The significance of free will*. New York: Oxford University Press.</bok>

<jrn>Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*, 556–567.</jrn>

<bok>Mele, A. (2006). *Free will and luck*. New York: Oxford University Press.</bok>

<bok>Mele, A. (2009). *Effective intentions*. New York: Oxford University Press.</bok>

<edb>Nahmias, E. (2011). Intuitions about free will, determinism, and bypassing. In R. Kane (Ed.), *The Oxford handbook of free will* (2nd ed., pp. 555–576). New York: Oxford University Press.</edb>

<edb>Nahmias, E., & Murray, D. (2010). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff, & J. Frankish (Eds.), *New waves in philosophy of action* (pp. 112–129). New York: Palgrave-Macmillan.</edb>

<jrn>Pereboom, D. (1995). Determinism *al dente*. *Noûs*, *29*, 21–45.</jrn>

<bok>Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.</bok>

<edb>Pereboom, D. (2012). Free will skepticism, blame, and obligation. In D. Justin Coates & N. Tognazzini (Eds.), *Blame: Its nature and norms*. (pp. 189–206) New York: Oxford University Press.</edb>

<other>Sartorio, C. (ms.). *The problem of determinism and free will is not the problem of determinism and free will*.</other>

<jrn>Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*, 543–545.</jrn>

<bok>van Inwagen, P. (1983). *An essay on free will*. Oxford: Oxford University Press.</bok>