

More Work for Hard Incompatibilism

Forthcoming. *Philosophy and Phenomenological Research*.

(Please do not cite without permission.)

1. Introduction

Hard Incompatibilism is the view that we lack the kind of free will that could make us truly deserving of blame or praise for our actions. The term ‘hard incompatibilism’ comes from Pereboom (2001) but versions of the position has been defended Voltaire, Diderot, Spinoza, Schopenhauer, Nietzsche, Clarence Darrow, Paul Edwards, Ted Honderich, Galen Strawson, Bruce Waller, Saul Smilansky, and Richard Double. Arguments for hard incompatibilism take roughly this form:

1. Some form of scientific naturalism (deterministic or indeterministic) is true.¹
2. The truth of some type of scientific naturalism entails that all actions we perform are the result of processes that trace back to factors beyond our control.
3. If an act is the result of processes that trace back to factors beyond the agent’s control, then the agent does not deserve blame or praise for that act. (The Transfer of Non-Responsibility (TNR) Principle.²)
4. Therefore no one can deserve praise or blame for any action whatsoever.

¹ With the possible exception of Ted Honderich, all contemporary skeptics about moral responsibility deny that we can be blameworthy or praiseworthy for our actions whether or not determinism is true. The term ‘hard incompatibilism’ alludes to what William James called the ‘hard determinism’ of Spinoza, Voltaire, Diderot, and Edwards (the view that determinism is incompatible with moral responsibility and that determinism is true). But hard incompatibilism is not committed to the truth of determinism.

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

Since most philosophers accept scientific naturalism of some sort, the debate focuses on premises (2) and (3). ‘Event-causal’ libertarians reject premise (2), arguing that physical indeterminism does grant us the type of control that can ground robust (desert-entailing) moral responsibility.³ Compatibilists typically reject premise (3), offering counterexamples to the ‘transfer of non-responsibility’ (TNR) principle.

I assume for this purposes of this paper that Pereboom’s “four-case-argument” successfully defends hard incompatibilism against both kinds of objections. Event-causal libertarians fail to show how physical indeterminism can provide the ultimate responsibility which they themselves believe is necessary for blameworthiness and praiseworthiness. And the compatibilist counterexamples fail to undermine the intuitive plausibility of the TNR principle expressed in premise (3).⁴ Even if we grant him these assumptions, however, Pereboom faces a further challenge. It is not enough to assert that premises (1) and (2) in the argument are true, and that premise (3) is highly intuitive. For hard incompatibilism to be vindicated, Pereboom must give *additional* reasons to reject the highly intuitive claim that people can be morally responsible for their actions *over* the highly intuitive TNR principle. Section 2 of this paper defends and develops this claim, and section 3 describes some ways that this additional challenge can eventually be met. The argument focuses on Pereboom specifically for the sake of clarity, but the argument applies to all positions that deny robust (desert-entailing) moral responsibility.⁵

³ See Kane (1996, 1999) and Ekstrom (2000) for two examples.

⁴ One might in addition deny that incompatibilist principles like TNR were intuitive even before compatibilist began hammering away at them. See Nahmias et al (forthcoming) for evidence that appears to undermine claims that people are natural or ‘pretheoretic’ incompatibilists.

⁵ Indeed, suitably revised, the argument can apply to all positions the debate, including compatibilist ones. Compatibilists, of course, face the challenge of showing that the TNR principle is *less* intuitive than the claim that people cannot be morally responsible for their behavior.

2. Hoisted on the Reflective Equilibrium Petard?

An example of the type of challenge I wish to develop is Pereboom's own objection to an argument by Ishtiyaque Haji. Haji's (1999, 2002) argument that determinism is incompatible with the existence of morally wrong actions can be expressed as follows:

1. Agent *S* has a moral obligation to perform [to refrain from performing] action *A* if and only if it is morally wrong for *S* to refrain from performing [to perform] *A*. The existence of obligations entails the existence of an ability to perform the obligation. **(OW)**
2. Agent *S* has a moral obligation to perform [to refrain from performing] action *A* (where *A* ranges over omissions as well) only if *S* has the ability to perform [refrain from performing] *A*. **(K)**.
3. Therefore it is morally wrong for *S* to perform [refrain from performing] *A* only if *S* has the ability to refrain from performing [perform] *A*. **(WAP)**
4. If determinism is true, then no one can have the ability to perform any act other than the act they actually perform.
5. Therefore determined worlds are devoid of wrong acts.⁶

There are several ways to attack this argument, but it is Pereboom's that interests me here.

Pereboom readily acknowledges the intuitive force of the key principles **(K)** and **(OW)** but writes:

The degree to which Haji's conclusions are unintuitive must be weighed against how unintuitive it is to reject one or more of his premises....If the components of the theory derived from these principles conform to our intuitions, that would provide theoretical support for them. But if such derived components do not conform to our intuitions, that would to some extent disconfirm these principles. I don't see how a principle's being an axiom in a moral theory would immunize it from such disconfirming pressures. (Pereboom, 2001, pp. 144-145).

⁶ This argument is drawn from Haji (2002), pp. 4-6. Pereboom (2001) discusses an earlier version of the same argument, Haji (1999).

According to Pereboom, the conclusion of Haji's argument—'nothing in determined worlds is right, wrong, and obligatory'—is so unintuitive that we should reexamine our initial acceptance of principles **K** and **OW**. To place further pressure on these principles, Pereboom shows that together with other premises to which Haji is committed, **K** and **OW** lead to two other counterintuitive results (**GR** and **BW**).⁷ Pereboom writes that were we to learn that determinism was true, "it seems arbitrary to privilege absolutely **K** and **OW** over **GR** and **BW**--and over the claim that judgments of moral obligation are sometimes true." (Pereboom, 2001, p. 146)

Pereboom does not indicate precisely how we are to resolve this battle of conflicting intuitions, but it seems clear that he favors a Rawlsian approach where we do our best to bring our intuitions and considered judgments on the matter into reflective equilibrium.

The problem for Pereboom is that the hard incompatibilist argument is vulnerable to the same line of attack that Pereboom employs against Haji. The unintuitiveness of the hard incompatibilist conclusion puts "disconfirming pressure" on the key incompatibilist premise—the TNR principle. My analogy is apt if the following two claims are true: (1) the hard incompatibilist conclusion is indeed unintuitive, and (2) that the TNR principle, like **K** and **OW**, is justified ultimately by an appeal to intuition. Claim (1) is uncontroversial. The belief that adult humans can sometimes deserve blame or praise for their behavior—call this the 'people can be morally responsible' or the 'PMR' belief—is acknowledged to be extremely intuitive even by those who conclude that it is false.⁸ Pereboom writes about Haji's conclusion that it "has the [unintuitive] consequence...that nothing Hitler ever did was wrong." No less hard to swallow is

⁷ GR: Sometimes actions that bring about the greatest good overall in worlds accessible to S are right for S. BW: Sometimes, when S is blameworthy for performing A, it was morally wrong for S to perform A.

⁸ Strawson (1986), a prominent skeptic about moral responsibility, believes that it is virtually impossible to accept this conclusion entirely—at least without undertaking a rigorous practice of meditation.

the hard incompatibilist conclusion that Hitler did not deserve blame or punishment for anything he ever did.

Claim (2) requires more support. First, it is worth noting that whether or not defenses of TNR *must* appeal to intuition, all contemporary incompatibilists *have* appealed to intuition in their defense of incompatibilist principles. Van Inwagen, for example, writes the following about principle beta—the ‘transfer of powerlessness’ principle that he later applies explicitly to moral responsibility as well:⁹

I must confess that my belief in the validity of Beta has only two sources, one incommunicable and the other inconclusive. The former source is what philosophers are pleased to call "intuition".... The latter source is the fact that I can think of no instances of Beta that have, or could possibly have, true premises and a false conclusion. (Van Inwagen, 1983, pp. 97-99)

Note that van Inwagen’s two “sources” are really quite similar, since someone with radically different intuitions from van Inwagen’s could presumably come up with counterexamples rather easily. (Anyone with the intuition that an agent in a particular case is morally responsible for a state of affairs he *clearly* could not have prevented from obtaining could simply use that case as a counterexample to van Inwagen’s TNR principle (rule B).¹⁰) And Pereboom, in defense of what he often calls the incompatibilist intuition, employs a ‘generalization strategy’ that works like this. First, he presents specific cases in which an agent is intuitively exempt from morally responsibility, and then he shows how those cases are (in the sense relevant to moral responsibility) identical to *all* cases involving human action. The generalization strategy cannot

⁹ Van Inwagen writes that rule B—the transfer principle relating to moral responsibility-- “perfectly parallels” Beta and in support of rule B refers the reader to his defense of Beta (p.187-188).

¹⁰ Consider a man, for example, who intuitively deemed himself blameworthy for his great-grandfather’s treatment of slaves. This intuition would be a counterexample to Rule B. Thanks to Eddy Nahmias and an anonymous referee for encouraging me to develop this point..

get off the ground, however, unless one shares with Pereboom the intuitions that the agents in initial cases are not morally responsible for their behavior.¹¹

Must incompatibilists appeal to intuition in their defense of the TNR principle? That is a more difficult question, one that I will not in this paper in any detail. It is not clear to me how one could even begin to develop a plausible theory of moral responsibility that disregarded intuition entirely. One might try to mount an argument that the concept of blameworthiness essentially involves the belief that the act in question cannot be traced to factors beyond one's control. But defending this claim in a non-question-begging manner would be difficult since it appears that people employ the concept of desert without committing themselves to beliefs about the originating causes of the behavior in question. (One interpretation of the doctrine of original sin, for example, is that human beings can deserve punishment for the sins of Adam and Eve. Whether or not one thinks this is a *justified* assignment of desert is irrelevant to the present discussion. The point is that moral responsibility likely does not *conceptually* involve anything like the TNR principle.¹²) A straightforwardly naturalistic defense of TNR would seem to run afoul of the Moorean/Humean arguments used against naturalistic justifications of objective moral values. This is because TNR is essentially a normative claim about the conditions under which it is *fair* to blame, praise, punish and reward people.¹³ Indeed, one might see TNR as a moral axiom, which is precisely what Haji calls his (Kantian) moral axioms **K** and **OW**. Of course, it would take significantly more analysis to conclude that it is *impossible* to defend the

¹¹ See Pereboom (2001), esp. Chapter Four. Of course, intuitive plausibility is no less relied upon in the cascade of counterexamples that compatibilists present in their attempt to undermine incompatibilist principles.

¹² Thanks to James Gibson for suggesting this example. The previous case (see note 10) can also support my claim here: the question 'do the great-grandchildren of slave-owners deserve any blame for their ancestors' treatment of slaves before the civil war?' seems to be an open one. We might find it intuitively obvious that they cannot deserve blame or praise for the actions of our ancestors (since they had no control over their ancestor's actions), but the question does not seem to involve a conceptual mistake

¹³ See Nichols and Vargas's "Reply to Levy", *Philosophy, Psychiatry, and Psychology*. Forthcoming. Vargas (2004) provides an excellent discussion of the various normative dimensions regarding judgments about moral responsibility.

TNR principle without appealing to intuitions. For now, however, it is enough to claim that no contemporary incompatibilist theory, including Pereboom's, has even made this attempt.

If it is true that Pereboom's conclusion is counterintuitive and that a key principle on which the conclusion relies appeals to intuition, then it seems that Pereboom's criticisms of Haji can be directed at his own position. What's sauce for the goose is sauce for the gander: if Haji's unintuitive conclusion places disconfirming pressure on the intuitive principles that lead him there, then Pereboom's unintuitive conclusion should place disconfirming pressure on the TNR principle. Pereboom writes:

It is not clear that a position that denies **GR**, **BW**, and that actions are sometimes morally obligatory, right, or wrong, while maintaining **K** and **OW** is superior to a view that, say, rejects **OW** and the claim that actions can be morally obligatory but accepts **GR**, **BW**, **K** and that actions are sometimes right or wrong. (Pereboom, 2001, p. 145.)

Similarly, one might write of Pereboom: 'it is not clear that a position that denies that people can be blameworthy or praiseworthy for anything is superior to one that accepts that human adults sometimes deserve blame for bad behavior but denies that agents can only deserve blame if their actions cannot be traced back to factors beyond their control.' It seems just as 'arbitrary' to 'privilege absolutely' the TNR belief over the PMR belief.¹⁴

Pereboom in the end does not claim to have decisively undermined Haji's argument, and notes that Haji's position "might win out in the end." But his analysis requires Haji to give additional reasons other than the intuitive plausibility of **K** and **OW** in support of his conclusion. Hard incompatibilists face the same requirement. They must do more than defend the intuitiveness of the TNR principle. They must demonstrate that the unintuitiveness of the hard

¹⁴ See also Nichols (2006) for a suggestion that reflective equilibrium is the most natural way to resolve the inconsistency in our intuitions about moral responsibility. Nichols presents very preliminary evidence that the conclusion may turn out to be a compatibilist one, but as he notes, much more work needs to be done before any side can claim even a partial victory.

incompatibilist conclusion does not place so much ‘disconfirming pressure’ on the TNR principle that we have reason to reject it in spite of its intuitive plausibility.¹⁵

3. Responding to the Challenge

Nichols (forthcoming) helpfully distinguishes among three dimensions of the free will debate: (1) the descriptive project, (2) the substantive project, and (3) the prescriptive project. The descriptive project is to determine the origins and the character of folk intuitions about moral responsibility. The substantive project is to determine whether the beliefs arising from these intuitions are correct. And the prescriptive project is (broadly speaking) to determine the ethical implications of what we learn from the substantive project. Skeptics about moral responsibility (myself included) have tended to keep the substantive and prescriptive projects distinct. First, we argue for *truth* of the claim that there is no such thing as moral responsibility. *Then* we examine the implications of this conclusion and address questions like: how would, or should, denying moral responsibility affect our metaethical beliefs? How would it affect our interpersonal relationships? How should it affect our approach to criminal justice? What are the implications for issues relating to the meaning of life? Hard incompatibilists like Pereboom tend to give rather optimistic answers to these questions, but the answers are not intended to have bearing on the truth of hard incompatibilism. As one author puts it: “Either we are capable of being robustly morally responsible for our behavior or we are not; metaphysical reality does not

¹⁵ Some may view Double (1998) as arguing for a similar conclusion. Double claims that one’s conclusion about whether we have free will depends on the ‘metaphilosophy’ that we embrace. I agree with Double that different metaphilosophies may produce different conclusions about free will and moral responsibility, but I also think that even two people who embrace a single metaphilosophy, including Double’s favored “Philosophy as Continuous With Science,” may legitimately arrive at different conclusions. What may make disagreement irresolvable is not that two people have different metaphilosophies, but rather that two people have fundamentally different starting intuitions about cases relating to moral responsibility.

tailor itself to our hopes and needs.”¹⁶ If what I argue in this paper is correct, however, an analysis of the implications (metaethical, ethical, practical) of hard incompatibilism has *direct bearing* on the truth, or plausibility, of the hard incompatibilist conclusion. In other words, the prescriptive project (and certainly the descriptive project) is an essential part of the substantive project. If the hard incompatibilist conclusion has unintuitive metaethical, ethical, or perhaps even pragmatic consequences, then we have more reason to reject TNR and preserve PMR (the belief that people can be morally responsible for their behavior). Since our intuitions ultimately ground TNR, it seems that the method of wide reflective equilibrium is the best way, perhaps the only way, to decide which belief—TNR or PMR—to accept. We cannot employ this method without consulting our feelings and intuitions about the implications of the hard incompatibilist conclusion.

Of course, reflective equilibrium may ultimately lead us to accept TNR and reject PMR—indeed, I believe it will, although I cannot argue for that claim here. I will conclude, however, by suggesting some factors that must be considered when deliberating over which of the two intuitive beliefs (TNR or PMR) we have more reason to reject.

1. Which belief is stronger? Which is more difficult psychologically to give up?
2. Which belief better coheres with other well justified beliefs, moral and non-moral?
3. Which belief can more plausibly be ‘explained away’—in other words, accounted for in such a way that does not require that the belief is true?
4. Which intuition or belief has more pragmatic value?

Answering these questions and relating them explicitly to the substantive arguments about moral responsibility is the additional ‘work’ I refer to in the title of this paper. The good news is that a much of this work has been done, although perhaps not with the goal of achieving reflective

¹⁶ Sommers, 2007, p. 342

equilibrium in mind. Experimental philosophers are at work trying to shed light on (1),¹⁷ and research in cultural anthropology, experimental economics, and social psychology has bearing on this question as well.¹⁸ Regarding (2), I have already noted that Pereboom presents an optimistic view of the ethical implications of hard incompatibilism. A vindication of this optimistic perspective would remove some of the ‘disconfirming pressure’ from the TNR principle.¹⁹ As for (3), several authors have argued that the belief in PMR can be ‘explained away’ with an account that incorporates both evolutionary and cultural factors.²⁰ If the same cannot be said for the TNR belief, then we arguably have more reason to accept it over the belief in PMR. Question (4) has yet to receive a comprehensive analysis, although a number of authors have addressed the practical day-to-day implications of denying moral responsibility, at least tangentially.²¹

So we do not need to start from scratch. But all of this work must be developed and formulated more carefully, and it must be related explicitly to the substantive question of whether or not we can be morally responsible. Until this is done, one may legitimately find every premise in the (valid) hard incompatibilist argument to be true or intuitively plausible yet still reject the hard incompatibilist conclusion.²²

Tamler Sommers

University of Minnesota, Morris

¹⁷ See Nahmias et al (forthcoming) and Nichols and Knobe (forthcoming) for representative examples.

¹⁸ The experiments in Henrich et al (2004), in which anthropologists run ultimatum and public goods games on subjects from 15 small scale societies has, I believe, deep relevance to the key question of whether intuitions about fairness and blameworthiness shared across cultures.

¹⁹ See also Sommers (2007a)

²⁰ See Sommers (2007b) Greene and Cohen (2004), and Nichols (2004).

²¹ See among others, Honderich (1983), Wolf (1981), Kane (1996), Sommers (2005), and Smilansky (2000).

²² Zimmerman (1987) rather playfully accuses Nagel in “Moral Luck” of accepting the premises in a valid argument while denying the conclusion. Perhaps Nagel had considerations like these in mind when he allowed himself to (seemingly) do this.

References

- Double, R. 1996. *Metaphilosophy and Free Will*. Oxford University Press.
- Ekstrom, L. 2000. *Free Will: A Philosophical Study*. Westview.
- Fischer, J. and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.
- Greene, J. and Cohen, J. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical Transactions of the Royal Society of London*. 359:1775-1778.
- Haji, I. 1999. "Moral Anchors and Control," *Canadian Journal of Philosophy*. 29:175-203.
- Haji, I. 2002. *Deontic Morality and Control*. Cambridge University Press.
- Henrich, J. et al. 2004. *Foundations of Human Sociality*. Oxford University Press.
- Honderich, T. 1993, *How Free Are You?* Oxford University Press.
- Kane, R. 1996. *The Significance of Free Will*. Oxford University Press.
- Kane, R. 1999. "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism." *Journal of Philosophy*. 96/5: 217-240
- Nagel, T. 1979. "Moral Luck." In *Mortal Questions*. Cambridge University Press.
- Nahmias, E. et al. "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research*. Forthcoming
- Nichols, S. 2004. "Folk Psychology of Free Will." *Mind & Language*, 19, 473-502.
- Nichols, S. 2006. "Folk Intuitions about Free will." *Journal of Cognition and Culture*, 6.
- Nichols, S. forthcoming. "How Can Psychology Contribute to the Free Will Debate?" In J. Baer, J. Kaufman, & R. Baumeister (eds.) *Psychology and Free Will*, Oxford University Press.
- Nichols, S. and Knobe, J. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Nous*. Forthcoming.
- Pereboom, D. 2001. *Living Without Free Will*. Cambridge University Press.
- Smilansky, S. 2000. *Free Will and Illusion*. Oxford: Oxford University Press.
- Sommers, T. 2005. "Beyond Freedom and Resentment: An Error Theory of Free Will and Moral Responsibility." Dissertation. Duke University.

- Sommers, T. 2007a. "The Objective Attitude." *The Philosophical Quarterly*. 57, 228: 321-342
- Sommers, T. 2007b. "The Illusion of Freedom Evolves," in *Distributed Cognition and the Will*, David Spurrett, Harold Kincaid, Don Ross, Lynn Stephens (eds). MIT Press.
- Strawson, G. 1986. *Freedom and Belief*. Oxford: Clarendon Press.
- Van Inwagen, P. 1983. *An Essay on Free Will*. Oxford University Press.
- Vargas, M. 2004. "Responsibility and the Aims of Theory: Strawson and Revisionism" *Pacific Philosophical Quarterly* 85:218-241.
- Wolf, S. 1981. "The Importance of Free Will." *Mind*, XC: 386-405.
- Zimmerman, M. 'Luck and Moral Responsibility.' *Ethics*, 97: 374-386