Freedom and Unpredictability

MICHAEL GARNETT

Birkbeck College, University of London

ABSTRACT *In* A Metaphysics for Freedom *(2012), Helen Steward proposes and defends a novel version of the libertarian account of free action. Amongst several objections that she considers to her view, one that looms particularly large is the* Challenge from Chance*: 'the most powerful, widely-promulgated and important line of anti-libertarian reasoning' (2012: 125). This paper begins by arguing that Steward's response to the Challenge (or, at least, to one strand of it) is not fully convincing. It then goes on to explore a further possible libertarian line of defence against the Challenge, arguing that it, too, ultimately fails. The conclusion is that the Challenge remains an important source of dialectical advantage for the compatibilist.*

I. Two Challenges from Chance

Steward characterises the Challenge from Chance as the view that 'the denial of determinism merely introduces an unhelpful randomness into the causal chains that underlie our intentional activity, and that such randomness could never help us to understand how free agency is possible' (2012: 125). Thus if free action is behaviour not entirely subject to deterministic causal laws, as the libertarian claims, then free action is, to that extent, random or chancy action, and it is difficult or impossible to see how such random behaviour could be meaningfully described as 'free' (or, indeed, even as 'action').

Yet what Steward calls 'the' Challenge from Chance is, I believe, most helpfully understood as a pair of related but logically independent problems. We may call these the *agency problem* and the *rational cost problem*. The former is the worry that, if what I end up doing is in some sense just a matter of luck, then there is no relevant sense in which what I end up doing is truly *up to me*. A good way of drawing out this concern is provided by Peter van Inwagen's 'rollback' case, in which Alice, facing a difficult decision between lying and truth-telling, freely chooses (in the libertarian, indeterministic sense) to tell the truth. Immediately after she does so, however, God reverts the universe to its exact state just a minute previously, and lets it run forward again. Since her 'second' decision, like her 'first' one, is undetermined by prior causes, there is no guarantee that she will do the same again. As van Inwagen puts it:

> Now let us suppose that God *a thousand times* caused the universe to revert to exactly the state it was in (and let us suppose that we are somehow suitably placed, metaphysically speaking, to observe the whole sequence of 'replays'). What would have happened? Well… sometimes Alice would have lied and sometimes she would have told the truth… Is it not true that as we watch the number of replays increase, we shall become convinced that what will happen on the next replay is a matter of chance? (2000: 14-15).

Thus it may come to seem that what Alice does on any given occasion is simply *up to chance*, and therefore not *up to her*. This is the first of the two challenges from chance.

The second is the problem that, if a free agent's processes of practical reasoning necessarily contain elements of randomness or chance, then a free agent must always be at risk of acting irrationally. Thus suppose that an agent confronts an opportunity set that provides just one rational option. Under determinism, such an agent might be so constituted as to be *guaranteed* to make the rational choice. Given some measure of indeterminism in action or deliberation, however, this cannot be the case. For the libertarian, it therefore seems, free agents are always at risk of irrationality. Yet this means that the freedom on which the libertarian insists is simply the freedom to be irrational, which is a freedom that, surely, we would be better off without. And it is implausible, other things equal, to suggest that free agency (or agency itself) depends crucially on our possession of a type of freedom that is worse than useless. As Susan Wolf argues, the freedom to be irrational is one that one could never have reason to exercise; nor, given this, could it be a freedom that one could intelligibly wish to have, since:

> Why should one want an ability that one never wants to exercise? Why should one care about being locked in a room—or, better, in a world—out of which one cannot *conceivably* want to go? Why should one mind if, to put it in extreme terms, one is *inescapably* sane? (1990: 57)

Thus such libertarian freedom cannot be of a variety 'worth wanting' (Dennett 1984).

The rational cost problem is independent of the agency problem, since even if we are convinced that whatever Alice chooses (in any given replay) is relevantly up to her, the libertarian still faces the problem of explaining how she could conceivably *care* about having the freedom to lie when she has better reason to tell the truth (or vice versa). And the agency problem is independent of the rational cost problem, since even were we persuaded of the value of being free to act irrationally, the libertarian would still need to explain how this could be a freedom for *us* to act irrationally (or for us to *act* irrationally). Both are potentially serious problems for the libertarian, and together they constitute the Challenges from Chance.

Although Steward does not explicitly distinguish the problems, she addresses both in detail. Thus, as regards the agency problem, she suggests that it is compelling only insofar as the libertarian has failed to provide any positive account of what it is for an action to be 'up to' an agent in the relevant sense; that is, that it is not (or, at least, that we are not simply entitled to assume that it is) merely indeterminism *in itself* that generates the worry (2012: 168-9). Steward goes on to supply precisely such a positive account (2012: 197-247), but I do not assess it here; for the purposes of this paper, I assume for the sake of argument that some such solution to the agency problem is possible, focusing instead on the rational cost problem. I argue that the difficulties it raises for the libertarian run deep.

II. Steward on the Rational Cost Problem

Steward's discussion suggests two lines of response to the rational cost problem. The first lies in her observation that, when an agent chooses to *φ*, the alternative possibility on which the libertarian must insist is not that the agent might have *chosen not to φ*, but simply that the agent might *not have chosen to φ* (2012: 155). Thus the libertarian need not claim, as many compatibilists seem to assume she must, that a free agent must have been able to do something despite having no reasons in favour of (or, indeed, having decisive reasons against) doing it. Instead, she must claim simply that a free agent must have been able to refrain from doing what she in fact did. As regards such mere powers of refrainment the rational cost problem is, it seems, much less severe.

To illustrate this, Steward considers a case in which Joe deliberates about whether or not to move in with his girlfriend, sees that he has excellent reasons for doing so, and accordingly decides to do so at *t*. She concedes that there is 'simply no coherent way of understanding' how Joe, aware of his plentiful reasons for moving in, could have decided not to move in; 'we can only conceive of the possibility of such a "decision" occurring, if we can conceive of it at all, as a kind of random upsurge of total irrationality into Joe's psychological life' (2012: 169-70). Yet, as she goes on to point out, this need be no implication of libertarianism. All the libertarian need claim is that Joe might not have decided *at t* to move in—and it is unlikely that Joe had any reason for making his decision just *then*.

What about the case in which he does have such a reason, for instance where his girlfriend has given him a time-limited ultimatum? Steward responds as follows:

> though not deciding at *t* to move in with his girlfriend would have been irrational in one way (because it prevents Joe from doing something he very much wants to do), it is not at all irrational in another. We have a general tendency, if we are prudent, not to rush into irrevocable decisions without careful thought and there therefore *are* reasons speaking for refrainment from deciding in the case imagined, because there are always general reasons speaking for caution and further thought (though of course, they can be outweighed by the need for urgency in a given case). (2012: 172)

Yet in cases where such reasons *are* outweighed, it seems the libertarian is still apt to find herself requiring that the agent in question possess the freedom to be irrational, and the rational cost problem remains. To take an even clearer case: suppose that Peter Singer has lost his mind and is credibly threatening to kill your family on the count of ten unless you press a button that will transfer ten pounds from your account to Oxfam; suppose further that you love your family, believe Oxfam to be a good cause, can easily spare ten pounds, and that Singer has already reached 'nine'. Of course, you press the button. But how are we to conceive of the possibility of your having refrained from acting just at that moment? Could this be anything other than 'a kind of random upsurge of total irrationality' into your psychological life?

Of course, even in this case there remain features of your action to be settled in the absence of decisive reasons: whether you push with your right or left hand, for instance, and the precise level of force with which you push (c.f. Steward 2012: 176-96). Yet the problem concerns the libertarian's handling of the claim that your action is also a settling of *whether you push*. To deny that agents can truly settle matters such as this, simply because their reasons incline clearly in one direction, is to treat reasons themselves as constraints on the scope of one's agency, a move which Steward decisively rejects (2012: 141-4). Yet to permit that they can is, it seems, to be vulnerable to the rational cost problem.[1]

Thus while Steward's focus on refrainment may help to show how libertarian freedom need not *always* amount to the freedom to be irrational, it falls short of providing a complete solution to the rational cost problem. Hence her second response. This is to point out that, since the metaphysical openness that entails the possibility of irrational action is, on her view, necessary for agency, any rational cost associated with it will be outweighed by the incalculably larger benefit of agency itself. Thus

> even if weakness of will is not useful or valuable to an agent, it might nevertheless be essential to the very *existence* of such an agent. For if, as I am arguing, an agent has to be a settler of matters at the time of action, it will need to be possible for her *not* to act, at any given moment, on a previously formed intention to φ… if freedom depends on agency (as it surely does) and if the metaphysical possibility of weakness of will is a necessary concomitant of the power of agency, the metaphysical possibility of weakness of will will be a necessary condition of freedom, notwithstanding what is, from another point of view, its uselessness *to* the agent whose existence makes it possible. (2012: 161)

[1] The same issue can be raised about Joe: he may settle the exact moment of his deciding to move in, but does he not *also* settle whether he decides to move in? And what could be our grounds for denying that he does, if we do not think that the mere decisiveness of one's reasons compromises one's agency?

That is, it is better to be a fallible agent than an infallible automaton: the rational cost pales against the agential benefit.

In assessing this second response we must keep in mind the broader dialectical situation. Recall that a proponent of a philosophical position has two tasks: first, that of elaborating a position that makes sense in its own terms, and, second, that of convincing others that the position is the best available. *Prima facie*, the rational cost problem causes trouble for the libertarian on both counts: it renders the position less satisfying for libertarians themselves, and it weakens it in relation to compatibilism.

Whereas Steward's second response succeeds in meeting the first of these challenges, however, it fails in meeting the second. This is because the response assumes the very libertarianism for which Steward is attempting to argue. No compatibilist, for instance, will accept that agency requires metaphysical openness. This is of course no problem at all when it comes to showing why libertarians need be *internally* untroubled by the rational cost problem. But it is problematic when it comes to swaying others. Imagine, if you will, that the libertarian and the compatibilist are debating before an audience of freewill agnostics. The compatibilist raises the rational cost problem: is it not implausible, she asks, to suppose that agency requires a freedom that is worse than useless? In reply, it will not do for the libertarian to point to the benefit yielded by this metaphysical freedom in making agency possible: this will carry no weight with the agnostics, and by the time she has persuaded them of this, she will already have persuaded them of libertarianism. In this dialectical context, therefore, Steward's second response is question-begging.

Neither of Steward's responses, then, succeeds in fully defusing the rational cost problem as a source of dialectical disadvantage for the libertarian. In the rest of this paper, I wish to consider an alternative possible line of response. Despite some initial promise, however, I ultimately argue that it, too, fails fully to solve the problem.

III. A Problem Shared

The alternative strategy I have in mind is that of *generalising* the rational cost problem in order to permit a *tu quoque* response to the compatibilist. Now, *tu quoque* responses are of course of no use when it comes to overcoming internal obstacles to one's position but, as we have seen, Steward's response already achieves this. When it comes to convincing others that one's position is the most compelling, by contrast, *tu quoque* responses are potentially effective, and it is with this second philosophical task that we are now concerned.

Put simply, the idea is this: the rational cost problem is not a problem for the libertarian alone, but for anyone who wishes to make room for any significant degree of *unpredictability* in their account of free agency. Moreover, compatibilism, if it is to be a plausible theory, must allow for unpredictability in some sense. The rational cost problem is therefore a shared one, and not a special source of dialectical disadvantage for the libertarian.

Allow me to elaborate. Ordinary people like you and I are not fully predictable to one another. Let me call this property, of being unpredictable to other ordinary agents, *ordinary unpredictability*. The existence of ordinary unpredictability is a truism accepted by all parties to the freewill debate. Nevertheless, it is one of which libertarians and compatibilists give distinct explanations. For the libertarian, our ordinary unpredictability is explained (at least in part) by our *metaphysical* unpredictability—an 'in principle' unpredictability entailed by the fundamental metaphysical openness of our actions. For the compatibilist, by contrast, our ordinary unpredictability is explained simply by our epistemic limitations: we are exceedingly complex systems and, while it could in principle be possible for some super-powered observer to predict fully what a human being will do, it is in practice absolutely impossible for you or I to do the same.

Now, there is a very deep and widespread intuition that the fact of our ordinary unpredictability bears some important connection to the fact of our free agency. It seems to be an important feature of dealing with free agents that, no matter how well you know them, you can never be *quite* sure just what they will do. Even one's closest friends are liable to surprise, in a way that strikes many people as somehow bound up with the idea that we are free agents and not mere automatons. Think, for instance, of the countless beginning philosophy students who, upon encountering the freewill debate, attempt to demonstrate their freedom by doing (or, more often, just affirming the possibility of their doing) something spontaneous and unpredictable; inapposite as this invariably is to the immediate matter at hand, it nevertheless expresses this same deeply-held feeling that the standing possibility of such unexpected behaviour has some important bearing on our freedom. Conversely, to imagine a being that is fully predictable to ordinary observers, the behaviour of which unfolds in accordance with simple and entirely transparent mechanistic principles, is to imagine a being that is *prima facie* lacking in (at least some important type of) freedom. This is the case with some (though certainly not all) of the lower animals: the point at which we can predict with certainty (or thereabouts) how a creature will behave is often also the point at which we lose our grip on the thought that it could possibly be a free agent, that there is 'anybody home'.[2]

Let me call the underlying thought here the *unpredictability intuition*. It is a vague intuition, to the effect that genuinely free agents must be, at least to some

[2] Dennett 1984: 13; see also his preceding discussion of the wasp *Sphex* (1984: 10-13).

significant extent, at least ordinarily unpredictable. This characterisation leaves entirely open not only what is the best explanation of our ordinary unpredictability (be that metaphysical or merely epistemic), but also on just what sense of 'freedom' it is rightly taken to bear (be that libertarian or compatibilist). Nevertheless, I take the unpredictability intuition—not least because of this very vagueness—to be one that it is reasonable to expect any plausible theory of free agency to find some way of accommodating; that is, it would be surprising to discover, and we would need a good argument to accept, that there is after all *no* sense in which it is true. Put in the terms of §2, it is an intuition that it is reasonable to assume would be prevalent amongst an audience of agnostics; and it would surely constitute a strike against a theory, in their eyes, were it unable to make any sense of it.

However, any theory that *does* so find a way of accommodating the unpredictability intuition will then find itself saddled with some version of the rational cost problem. This is because our ordinary unpredictability is, to at least some extent, dependent upon our liability to irrationality. To be sure, it is not *wholly* dependent upon our liability to irrationality: there are many cases in which reason does not prescribe a unique course of action, as well as cases in which we are simply ignorant of one another's reasons (though the compatibilist cannot afford to put too much emphasis on this latter observation, as will be explained below). Nevertheless, were we all both perfectly and unavoidably rational, we would be far more predictable to one another than we currently are. Being perfectly rational we would always understand the requirements of reason and, additionally, we would each know that there were absolutely no chance at all of the other failing to act rationally. We may of course dispute just *how* mutually predictable perfectly rational agents would be—this will no doubt depend upon our theory of rationality. But on many theories they would likely be predictable enough as to seem, by the lights of the unpredictability intuition, at least somewhat deficient in some valuable type of freedom; in Daniel Dennett's words, each would risk being 'bereft of *personality*, a mere conduit for Truth or Doing the Right Thing, not a unique and idiosyncratic actor on the world stage' (1984: 70). Thus ordinary unpredictability requires at least some possibility of irrationality: ordinary unpredictability carries a rational cost. And if free agents must be ordinarily unpredictable—that is, if the unpredictability intuition is in any sense true—then free agency *itself* carries a rational cost. The rational cost problem is a general one.

We therefore have the following argument:

(1) Our ordinary unpredictability is in some important way bound up with our status as free agents (the unpredictability intuition).

(2) At least some significant measure of liability to irrationality is implicated in our ordinary unpredictability.

(3) *Therefore*, at least some significant measure of liability to irrationality is implicated in our status as free agents: the rational cost problem should be a problem for all accounts of free agency.

However, this is an argument that the compatibilist will naturally attempt to resist. First, she may target (2). After all, much of what we have reason to do is dependent upon our tastes and desires, and we are often ignorant of one another's tastes and desires. To take a simple example: I cannot predict what you will order from a restaurant menu if I have no idea what kind of food you like, even if I know you to be perfectly rational, since what you have reason to order depends on what you like. Even perfectly rational agents, then, will be mutually unpredictable to a significant extent insofar as they lack knowledge of one another's tastes and desires. So (2), the compatibilist may conclude, is relevantly false, and the argument unsound.[3]

However, this kind of response to (2), while no doubt correct, is of little help to the compatibilist. For if it is our *ignorance* that is thus implicated in our ordinary unpredictability, and if our ordinary unpredictability is in some way bound up with our status as free agents, then a parallel version of this argument will show that our ignorance is in some way bound up with our status as free agents—a conclusion at least as unpalatable as (3). That is, this way of rejecting (2) simply threatens to saddle the compatibilist with an *epistemic cost problem* at least as problematic as the rational cost problem that she is attempting to avoid. The libertarian may therefore offer the compatibilist a choice: either join the libertarian in accepting the counterintuitive conclusion that free agency itself requires a liability to irrationality, or else adopt the equally counterintuitive conclusion that free agency requires ignorance (or a measure of both).

Given this, the compatibilist may instead be tempted to reject (1). This is especially likely given that the compatibilist may anticipate being unable to accommodate *any* version of the unpredictability intuition. As we have seen, compatibilists have no problem accommodating the fact of our ordinary unpredictability: we are highly complex and epistemically limited. But they have a *prima facie* problem explaining why this should have any bearing on our freedom. After all, why should *your* epistemic limitations have anything to do with *my* freedom? Moreover, this is a question to which the libertarian has a good answer,

[3] Thanks to Karin Boxer for pressing me on this point.

this being that your inability to predict my actions is simply a symptom of their metaphysical openness, which is in turn a necessary condition of their freedom. Obviously, the compatibilist cannot go down that route. So she may prefer to head the unpredictability intuition off at the pass.

Yet matters are not so simple. Recall that the unpredictability intuition has here been carefully characterised so as to avoid begging any questions against the compatibilist. It states a vaguely but deeply held feeling, widespread amongst those innocent of the philosophical debates, which theories of free agency may reasonably be expected to accommodate. Denying it outright carries a significant intuitive cost. So the libertarian may again present a dilemma: either the compatibilist must find some way of accommodating the unpredictability intuition, in which case she is herself subject to a form of the rational cost problem and so cannot employ it in her dispute with the libertarian, or she must reject the unpredictability intuition, thereby accepting a dialectical loss unlikely to be fully counterbalanced by the associated gain of freeing herself to raise the rational cost problem against the libertarian.[4] Either way, the libertarian has successfully defused the rational cost problem as a source of dialectical advantage for the compatibilist.

However, I shall now show that this line of argument is ultimately unsuccessful. There is indeed a means for the compatibilist to accommodate a version of the unpredictability intuition without thereby succumbing to the full force of the rational cost problem. Showing how the compatibilist might thus successfully grasp the first horn of this dilemma is the task of the remainder of the paper.

[4] Of course, denying the unpredictability intuition need not constitute an *internal* worry for compatibilism, just as the rational cost problem need not constitute an internal worry for libertarianism. For an audience of agnostics not yet convinced of either view, however, both plausibly represent strikes against their respective sides.

IV. Unpredictability for Compatibilists

How might mere epistemic unpredictability enhance an agent's freedom? Since epistemic unpredictability—that is, unpredictability to other agents—is an essentially *social* or *relational* notion, we might profitably attempt to answer this question by considering freedom in *its* social or relational forms. Indeed, the way for the compatibilist to accommodate the unpredictability intuition is simply to shift her focus from the metaphysical to the social.

Allow me then to outline the type of social freedom I have in mind. On the well-known 'negative' conception, freedom consists in the absence of (certain types of) interference by other agents. On the recently revived though long-standing 'republican' conception, by contrast, freedom consists in *immunity* or *resilience* to (certain types of) interference by other agents (Pettit 1997; Skinner 1998). On this latter view, a free agent is one that is resistant to subjection by foreign wills, and so difficult for others to manipulate or to control. It is with this latter, republican conception that the link with unpredictability may be discerned.

Resistance to subjection by foreign wills is a dispositional property of agents. Moreover, it is a property that is *conferred* on agents by their possession of certain base properties (just as, say, the property of *being a sedative* is conferred on a substance by its possession of certain base properties, such as that of *being a barbiturate* or *being an alcohol*). For example, if you are the subject of a legally enforced right not to be physically attacked (and thus relatively immune to threats of physical violence), you are to that extent difficult for others to control. Similarly, if you have the capacity to reason critically, such as to render you relatively immune to manipulation by sophistical argument, you are to that extent difficult for others to control. It is one's possession of these sorts of base properties (being a right-holder, having a capacity for critical rationality) that confers on one the higher-order dispositional property of being resistant to subjection to foreign wills.

Elsewhere I have argued in detail that possession of legally protected status, a capacity for critical reflection, and a healthy sense of one's own self-worth are all significant conferrers of resistance to interpersonal subjection, and hence of social freedom in its broadly republican sense (Garnett 2013). Now I wish to suggest that epistemic unpredictability is, in the same way, an important conferrer of such resistance. To see how, consider the following case from Derek Parfit:

> *Schelling's Answer to Armed Robbery*. A man breaks into my house. He hears me calling the police. But, since the nearest town is far away, the police cannot arrive in less than fifteen minutes. The man orders me to open the safe in which I hoard my gold. He threatens that, unless he gets the gold in the next five minutes, he will start shooting my children, one by one... I am in a desperate position. Fortunately, I remember reading Schelling's *The Strategy of Conflict.* I also have a special drug, conveniently at hand. This drug causes one to be, for a brief period, very irrational. Before the man can stop me, I reach for the bottle and drink. Within a few seconds, it becomes apparent that I am crazy. Reeling about the room, I say to the man: 'Go ahead. I love my children. So please kill them.' The man tries to get the gold by torturing me. I cry out: 'This is agony. So please go on.' Given the state I am in, the man is now powerless. He can do nothing that will induce me to open the safe. Threats and torture cannot force concessions from someone who is so irrational. (1984: 12-13)

Note that it is Parfit's unpredictability, and not his irrationality *per se*, that renders him uncontrollable. The robber's problem is that he no longer knows how to induce Parfit to act as he wishes: it may be, for all he knows, that an offer to sing the score of

Cats would result in Parfit opening his safe. So were the robber to find a manual detailing the exact changes the drug has rendered to Parfit's processes of practical reasoning, he could simply look up 'safe-opening' in the index of outputs and set about producing in Parfit the required input. Parfit, though still irrational, would then be predictable, and hence controllable.

Though this is an extreme case, it carries an important lesson. If I am to control your behaviour, I must be able to predict how you will respond to various stimuli. To the extent to which I cannot make such predictions, you are resistant to my control and manipulation. Thus unpredictability helps confer social freedom.[5]

This may feel like a familiar point. Dennett, for instance, in a section of *Elbow Room* titled 'The Uses of Disorder', argues that, since the social environment I inhabit may contain other agents that are potentially hostile to me, 'I have a reason, a meta-level reason, for wanting my mind to be unreadable, and this might well require that I avoid putting patterns into certain of my activities. The only way of assuring that there is no readable pattern in those activities is to make them random' (1984: 66-7). Moreover, he argues, such randomness and unpredictability is evolutionarily advantageous (p. 66); it is also epistemically advantageous, helping us to sample large domains, and practically advantageous, helping us to cut short potentially endless

[5] Must unpredictability confer social freedom (in this broadly republican sense) in every possible circumstance? Yes: since resistance to interpersonal control is a dispositional property, what matters is only how difficult one *would* be to control, *were* someone to attempt it (and not, for instance, whether anyone does in fact attempt it). Thus one may possess (or lack) social freedom even in the absence of potential manipulators or controllers. But must social freedom, and hence a measure of unpredictability, be *valuable* in every possible circumstance? No: in a world without potential controllers, we would surely have little reason to value freedom in this sense. To explain the unpredictability intuition, however, the compatibilist need only demonstrate a conceptual link between unpredictability and a type of freedom that we do in fact value. (Thanks to Anton Ford for pressing me on this point.)

deliberations (pp. 68-9). Yet despite this list of reasons for valuing randomness and unpredictability, Dennett fails to explain why any of it should have anything to do with *freedom*. Compatibilists can of course recognise all of these advantages, but their problem lies in explaining why an agent that lacks these advantages, that is wholly mechanistic and predictable in its behaviour, is in any way deficient in freedom specifically. In short, Dennett fails to show how the compatibilist can accommodate the unpredictability intuition.

For this we require the republican conception of freedom. Thus to be fully predictable is to be vulnerable to the domination of others, and to be vulnerable to the domination of others is to be (in at least one important sense) unfree. So whereas the libertarian is able to accommodate the unpredictability intuition by linking our *in-principle* unpredictability with the idea of metaphysical freedom, the compatibilist is able to do so by linking our *in-practice* unpredictability with the idea of social freedom. Compatibilism is thereby strengthened by its ability to match the libertarian in vindicating a version of this fundamental intuition.

V. The Rational Cost Problem Solved

In doing so, however, the compatibilist seems to open herself up to the rational cost problem. Unpredictability, we have seen, helps to confer uncontrollability, which is a form of social freedom. Moreover, the relationship is linear: the more unpredictable an agent, the more uncontrollable. And, at least for high degrees of unpredictability—and therefore for high degrees of uncontrollability—actual or likely irrationality is likely necessary. Thus Parfit renders himself maximally uncontrollable, by rendering himself maximally unpredictable, by rendering himself maximally irrational. If social freedom requires anything like the kind of extreme irrationality manifested by Crazy Parfit, we will surely feel that we are better off without it.

Of course, *some* degree of unpredictability is attainable without any possibility of irrationality: as previously noted, not every practical problem has just one rational solution, and we are often ignorant of one another's reasons. But for the ordinary levels of unpredictability that are intuitively associated with free agency, it may still be urged that at least some measure of liability to irrational action is necessary. So, having now accepted that free agents are (to some extent) unpredictable agents, the compatibilist opens herself to the complaint that free agents are therefore (to some extent) fallible and potentially irrational agents. This means that the compatibilist can seemingly no longer raise the rational cost problem as an objection to libertarianism without inviting a *tu quoque* response.

However, the compatibilist, unlike the libertarian, has the resources with which to solve her version of the rational cost problem. Indeed, she has two complementary responses available to her.

First of all, the overall rational cost faced by the compatibilist is likely lower than that faced by the libertarian, owing to a structural difference between the two accounts of unpredictability. This is because, whereas the libertarian links unpredictability to a notion of (metaphysically) free *action*, the compatibilist links unpredictability to a notion of (socially) free *agency*. The libertarian is therefore committed to the claim that an action is free (or that a piece of behaviour is an action) only if it is (metaphysically) unpredictable. This is a strong claim, since it requires that each and every action must be unpredictable; and, in any case in which the reasons incline clearly in one direction, it entails that the agent was liable to act irrationally. Indeed, the libertarian has no respite from this conclusion; it applies not only in the clear cases, such as Joe's decision concerning whether to move in with his girlfriend, but also in the very clearest cases, like that of Crazy Singer. By contrast, the compatibilist is committed to the different claim that an *agent* is free only if she is to some extent (epistemically) unpredictable. This is a weaker claim, insofar as it treats unpredictability as a global property of agents and not as a local property of their actions. In particular, it need not entail that a free agent be unpredictable with respect to every action. Instead, it may merely require that there be some threshold of global unpredictability below which she does not fall. Thus one need not be as unpredictable as Crazy Parfit to meet the relevant requirement; nor must one be even slightly unpredictable on every conceivable occasion, even when faced with Crazy Singer. So although the compatibilist may have to concede that freedom requires a general liability to irrationality, she need not accept as extreme and austere a version of this idea as that to which the libertarian appears committed.

Second of all, the liability to irrationality that the compatibilist must still concede may be shown to be worth the cost. That is, the compatibilist may argue that, *up to a point*, the possibility of irrationality is a price worth paying for the benefit of increased social freedom: the rationality cost is outweighed by the freedom benefit. Note that this parallels Steward's libertarian response, discussed in §2, of claiming that the possibility of irrationality is a price worth paying for the benefit of metaphysical freedom (this being necessary for agency itself). That response was rejected on the grounds that it is question-begging in the current dialectical context, since the compatibilist, in denying that metaphysical freedom is necessary for agency, sees no benefit in metaphysical freedom. The equivalent compatibilist response that we are now considering, however, is not similarly question-begging. This is due to the underlying asymmetry between libertarianism and compatibilism: whereas the compatibilist *rejects* libertarian freedom, arguing that we need concern ourselves *only* with compatibilist freedoms, the libertarian does not reject but typically accepts the importance of the compatibilist freedoms, arguing instead that we need concern ourselves *also* with libertarian freedom. Thus the libertarian may be expected to join the compatibilist in recognising the value of social freedom. Appealing to the value of this social freedom in attempting to meet (the compatibilist's version of) the rational cost problem is therefore not question-begging against the libertarian.

For these reasons, I conclude that the rational cost problem remains a source of dialectical advantage for the compatibilist. As we saw, Steward's response, although effective at explaining why libertarianism need not be troubled by the problem within its own terms, fails to neutralise the problem in the context of her dispute with the compatibilist. Moreover, the alternative, generalising strategy considered in §3 has ultimately proven of limited effectiveness; for we have seen that not only does the compatibilist have a way of making sense of the intuitive relationship between unpredictability and freedom, and hence of making an incursion into what is traditionally libertarian territory, but that she is able to do so without herself falling victim to the rational cost problem.[6]

References

Dennett, Daniel. C. *Elbow Room: The varieties of free will worth wanting*. Cambridge, MA.,: The MIT Press, 1984.

[6] Thank you to the organiser, Jennifer Horsnby, and to the other participants of the April 2013 *Workshop on 'A Metaphysics for Freedom'* at the Centre for the Study of Mind in Nature at the University of Oslo, especially to Helen Steward both for writing her excellent book and for her helpful comments on this paper. Thank you also to my wife, Syeda Khanom, for taking on a disproportionate share of our childcare responsibilities in order to enable me to meet the deadlines associated with completion of the paper.

Garnett, Michael. The autonomous life: a pure social view. *Australasian Journal of Philosophy* 9/14 (2013).

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Pettit, Philip. *Republicanism: A theory of freedom and government*. Oxford: Oxford University Press, 1997.

Skinner, Quentin. *Liberty Before Liberalism*. Cambridge: Cambridge University Press, 1998.

Steward, Helen. *A Metaphysics for Freedom*. Oxford: Oxford University Press, 2012.

van Inwagen, Peter. ‘Free will remains a mystery’. *Nous* 34, Suppl. 14 (2000), pp. 1-19.

Wolf, Susan. *Freedom Within Reason*. New York: Oxford University Press, 1990.