Précis of *A Metaphysics For Freedom*

Helen Steward

A Metaphysics for Freedom (2012) is a book that grew out of a certain frustration on my part with the state of the free will literature as it stood in the early years of the current century. For all its richness and ingenuity, this literature frequently failed to address what I regarded as crucial, foundational issues in the metaphysics of causation, the philosophy of mind, and the philosophy of action, issues which I was convinced mattered enormously to the debate. It had become usual to address the free will problem largely from an ethical perspective, and the links to discussions about moral responsibility, the justification of punishment, etc., had accordingly become very tight. In itself, of course, there is absolutely nothing wrong with treating free will as a problem in ethics, for it does indeed require to be considered from this perspective, but if one approaches the issue *solely* from this direction, certain ontological and metaphysical questions tend to be glossed over and thereby to become obscured. In particular, I felt convinced that the idea of agency *itself* needed a lot more scrutiny than it had tended to be given, and that it was a more distinctive and more complex concept than had been generally recognized. In order to reflect this conviction, I opened the book with an epigraph from Jean-Paul Sartre. Sartre writes:

> It is strange that philosophers have been able to argue endlessly about determinism and free-will, to cite examples in favour of one or the other thesis, without ever attempting first to make explicit the structures contained in the very idea of action. (1958, 433)

This idea—that there are what Sartre calls 'structures' contained in the very idea of action and that we need to get clearer about what exactly those structures are before we can embark properly on discussions about free will and determinism—was one that strongly motivated the book project. In particular, I began to think that at least some respectable motivations toward incompatibilism might derive not so much from ideas specifically about so-called *free* agency, a special power thought of as unique to human beings, mainly exercised in situations of profound significance, moral choice, etc., but rather from our deeply-rooted conceptions of something much more basic—agency itself, thought of as a power common to a wide range of animals.

When I thought about the world as I felt constrained to imagine it under the hypothesis of determinism, it wasn't merely human choices that seemed not to fit with the resultant picture, but also the activities of a whole

Res Philosophica, *Vol. 91, No. 3, July 2014, pp. 513–518*
http://dx.doi.org/10.11612/resphil.2014.91.3.13

range of scurrying, wandering, flapping and swimming creatures which I found myself strongly inclined to think of as determining for themselves, moment to moment, where precisely they would go and what exactly they would do next, whose various forays into and meanderings around the world I had great trouble thinking of as fixed in all their glorious detail by initial conditions at the so-called 'beginning' of the Universe, on the one hand, and the laws, on the other. Surely it is only as the individual animal moves, I thought, or at best very shortly before, as the nervous activity preparatory to those movements is initiated in the animal's brain, that it becomes metaphysically settled that these very motions are going to occur. I found I just could not bring myself to believe in the in-principle derivability of these very motions from initial conditions and laws. This view, as I put it in the book, is quite literally incredible. And so gradually, over the course of a few years, I developed the position that rather late in the day I decided to call 'Agency Incompatibilism'—the idea that agency itself is inconsistent with determinism, that there is something about the structure of the concept of action which makes it inimical to the idea that actions might be among the things which have deterministic causes.

Why might one think that there are structures contained in the very idea of action which are inconsistent with the thought that actions are determined events? My basic thought was a simple one—that we regard the actions and activities of animals, unlike events occurring in inanimate objects, as having their ultimate source with the agent of the action, in a way to which nothing corresponds in the case of non-agents. Substances which are not agents can of course cause things, but they are usually caused to cause the things they cause. A brick breaks a window—but that is because it was thrown at the window at high speed. A building collapses and kills someone—but it was caused to collapse by an earthquake which occurred just beforehand. Chains of causation in the inanimate world thus proceed back—potentially deterministically—beyond the substance itself, to the events which triggered the manifestation of its powers to produce further sorts of effect, and then in turn to the events which triggered those further events. But I felt convinced that this was not in general how we think pre-philosophically about the causation of events by *agents*. The causal chains which are explanatory of the occurrence of actions can indeed proceed back into the past beyond the action, but when this occurs, the mode of the relevant causation, generally speaking, is *influence*, not determination. It remains the agent herself who has to convert this influence into activity—and the details of this conversion are always at her discretion. And for this sourcing of things in the agent herself really to *mean* something, it seemed to me, for it to be anything over and above the mere power to cause that is shared by all sorts of other substances, agency would have itself to be a power inconsistent with determinism. Nothing is a truly ultimate source of events which produces those events only because it is wholly caused to produce them by something else—for then the ultimate source is just

the *other* thing. Whereas a true agent ought to have the power, as I put it, genuinely to *settle* with her actions for at least some propositions *p*, whether or not it will be the case that *p*. It has become common in recent years to distinguish two conditions often appealed to in the libertarian literature: the 'leeway' condition, which insists that a free agent must have alternate possibilities available, and the 'source condition', which insists that an agent must be an ultimate arché or origin, in some sense, of the events she produces as agent (Kane 1996; Pereboom 2001, 2003). But for me, the two conditions are connected together. My intuition was that one could only be a true source if one had leeway. Being a proper source of some state of affairs, as opposed to an inadequate sort of *ersatz*, requires that it is oneself who determines the matter—and that requires that nothing *else* does—that one is not determined to determine as one does. That, I felt, was a libertarian intuition worth inspecting in more detail for a distinctive variety of incompatibilism.

The compatibilist, of course, will tend to think that she has plenty of places to turn in order to respect what is right about both the sourcehood and the leeway intuitions without conceding anything to indeterminism. She might say, for instance, that it is when an event originates with something like a choice or a decision—or perhaps when it issues from an intention—that the source of an event can be regarded as being the agent herself. But I did not see how this could be the right answer. One worry which made me wary of this compatibilist idea was that choices and decisions are comparatively rare phenomena when it comes to agency. We do a great many of the things we do without choosing or deciding, and sometimes even without intending to do them, and this seems even more likely to be true of other, simpler creatures. I wanted to put into the spotlight the huge variety of actions we do unthinkingly, habitually, unconsciously, without reason, without planning or forethought, without deliberation, and which yet are our actions nonetheless. I wanted to say: even these actions seem *done*, they have their source in their agents. Causation by the mental, then, cannot be the key to understanding what is agential about agency, because so much agency proceeds without it. Moreover, although I think it is true that only agents can truly choose and decide things, intend things, and so on, to try to analyze what it is to act in terms of these other concepts of choice, intention, and the like seemed to me to be to get things the wrong way round. For one has to code something as an agent before the question whether it can choose or decide anything can even arise.

I mentioned above that the label 'Agency Incompatibilism' was one that I came to attach to my position only relatively late in the day. I am keen to point this out because it enables me to draw attention to something that I wouldn't want to have missed about the book. Labelling a view with this word 'incompatibilism' unfortunately instantly lines it up on one side of what is always perceived to be the major ideological divide in the free will area. And there can be something a bit tribal about the way

philosophers line up along these major fault lines in the subject, to the extent that one gets a serious hearing, sometimes, only from those who think of themselves as being on the same side of the line. But although the view argued for in *A Metaphysics* is a version of incompatibilism, I had always conceived of it, while writing the book, as a position that permitted major concessions to compatibilism on some very important issues; in a way, indeed, while writing most of the book, I thought of my position as a sort of middle way between compatibilism and traditional versions of incompatibilism. In particular, I thought that relocating the main source of respectable incompatibilist intuitions in the concept of agency itself would mean that it did *not* have to be located in some of the more implausible places where traditional incompatibilists had tried to position it, and *that* enabled me to take the compatibilist side in a number of crucial disputes. I did not, for example, have to require that in a situation in which an agent deliberates about whether to do *A* or *B*, and decides on the basis of good reasons or strong preferences, or whatever, to do *A*, it must have been physically or metaphysically possible that she choose to do *B* instead, even while holding fixed all antecedent factors, including the agent's motivations, desires, reasons, etc. It seemed to me that an agent might well have preferences, personality traits, and so on, which would make it utterly inconceivable that they would choose *B*, under such circumstances (i.e., without envisaging some change in the structure of their antecedent desires or beliefs). Moreover, I was free to agree with the compatibilist in insisting that the question whether an agent is free to do *A* or *B*, or has the power to choose *A* or *B*, is not the same question as the question whether, given the laws and initial conditions as fixed, the world might then proceed in such a way that *A* is chosen, and might also proceed in such a way that *B* is chosen. All that is required for Agency Incompatibilism is that no individual action be a necessitated event. And it does not follow, of course, from the fact that action *a* is not necessitated that action *b* must be metaphysically possible. All that is required is that it have been possible, consistently with the prior conditions and the laws, for *a* not to occur—a much weaker and less stringent condition. Moreover, the condition, note, is one that only relates to *a* as an individual action—it is perfectly consistent with the view that individual action *a* was not necessitated that there be no possibility, given laws and prior conditions, that some action of one or more of the *types* instantiated by *a* (say, type *A*) not occur. In its general lineaments, the course of reality over a given period might well be dictated by laws which make the non-occurrence of an *A*-type action within a certain time period psychologically or physically impossible. But what is not conceivable, I argued, is that something that is genuinely an action be a wholly necessitated event. For that would be inconsistent with the idea of the agent him/herself as the source of the movement or change the production of which constitutes the action.

Another compatibilist thought that I wanted to try somehow to accommodate was the insight that the postulation of microphysical indeterminism, in and of itself, does nothing to dispel the serious worries about how on earth the power of agency is possible. If it is hard to see how agency is consistent with determinism, it is equally hard to see how the postulation of mere microphysical indeterminism could allow for it. For there are at least two kinds of determinism that create worries for agency—one is the determinism which envisions the antecedent determination of what is present and future by what is past; but the other is a determinism which envisions the bottom-up determination of the activities going on within a large and complex system by the activities going on in its parts. Both these forms of determinism need addressing and the introduction of microphysical indeterminism, in and of itself, addresses only the first. A fully satisfactory account of agency has to address the second, also, and I had ambitions in the book to try to say something about how the bottom-up picture might perhaps be challenged. To address this second issue, I needed to develop, so I thought, a workable notion of top-down causation which might make the idea that whole substances are able, under certain circumstances, to wreak effects on their own parts seem more intelligible than traditional metaphysical outlooks tend to permit it to seem.

On top-down causation, the challenge, as I saw it, was to try to see how a large and complex entity, like an animal, might have any efficacy that did not just reduce entirely to the efficacy of its parts. For such a reduction, I feared, would mean the loss of the crucial property I was convinced agents must have—the property of being true settlers of things, ultimate sources of what occurs, not just substantial causes. In a mechanism that is working well, a part must do what its circumstances dictate it will do, and if it does not, well, that is just so much the worse for the mechanism. But if human beings are ultimately merely very beautifully constructed mechanisms, if all the parts of a human being do what they do merely in necessitated response to the prior activity of adjacent parts, then I feared that the phenomenon of human agency would disappear once more, this time succumbing to the relentless pressure of the bottom-up picture of the world that has become so very prevalent.

Much of the final chapter of *A Metaphysics* is therefore devoted to the attempt to see whether there might be any mileage in the concept of top-down causation, understood as the capacity for a whole to affect its parts in such a way that those effects do not just reduce down to the impact of parts on part. I tried to argue for ways of resisting the bottom-up picture, based on challenges to the idea that supervenience, in and of itself, implies that the evolution of reality *over time* is due entirely to the interactions of small parts, and also on some reflections on the concept of coincidence, and the impossibility of deriving from lower level laws and descriptions themselves, any understanding of how the incredibly complex coordination and ordering of lower level phenomena that is required in order for a

complex event such as a human action to occur, can be achieved. To understand that, I suggested, we might need the idea of the agent herself as top-down coordinator and organizer of some of the wanted collocations. I would be the first to admit, though, that this last chapter is very far from being the last word on the matter of top-down causation; much more thought needs to be given to the concept, and I think science as well as philosophy will need to contribute to any intellectually satisfactory picture.

More than anything, though, I wanted *A Metaphysics* to present not a fully worked out version of Agency Incompatibilism so much as some reasons for thinking that there was a position in this area of logical space that was worth working out. The idea that a serious form of libertarianism might be based on thoughts not merely about human creatures and their special powers, but rather on ideas about what animality in general might bring to the world seemed to me to be a project worth pursuing. This way of thinking about the issues promised to bring the prospects of a truly naturalistic libertarianism much closer, a libertarianism in which freedom would be found to emerge not from such evolutionarily recent phenomena as rationality or ethics, but from the deep and still largely mysterious foundations of biology, where much richer metaphysical resources might be available to be mined.

Helen Steward
E-mail : h.steward@leeds.ac.uk

References:

Kane, Robert. 1996. *The Significance of Free Will*. Oxford: Oxford University Press.
Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.
Pereboom, Derk. 2003. "Source Incompatibilism and Alternative Possibilities." In *Moral Responsibility and Alternative Possibilities*, edited by Michael McKenna and David Widerker. Aldershot: Ashgate.
Sartre, Jean-Paul. 1958. *Being and Nothingness*. Translated by Hazel E. Barnes. London: Methuen.
Steward, Helen. 2012. *A Metaphysics for Freedom*. Oxford: Oxford University Press.