Strategies for free will compatibilists

JOHN O'LEARY-HAWTHORNE & PHILIP PETTIT

While most contemporary philosophers probably take a compatibilist position on the free will issue, few bother to investigate the varieties of compatibilism available and to consider their respective attractions. Discussions of free will have been focused sharply on the issue between compatibilism and incompatibilism: the issue as to whether free will is consistent or inconsistent with a picture of the universe under which everything happens out of mechanistic necessity. Such a focus is perfectly understandable but it has tended to put other relevant questions out of sight, in particular questions that bear on the forms of compatibilism available, and their respective merits.

This essay is an attempt to redress the balance of attention and to offer an overview of the different broad strategies that compatibilists may pursue. We start from a formulaic account of what it is for an agent to do something freely which all or nearly all compatibilists can be expected to endorse. We go on to look, then, at the different ways in which this account may be interpreted by different compatibilists. And we offer a concluding remark on the nature of the contest between the different strategies.

1. *Formulaic compatibilism*

Our formulaic account of what it is for someone to act freely – to act freely by compatibilist lights – is this:

> X chooses freely to φ if and only if the relevant antecedents of the choice leave the φ-ing up to X.

All compatibilists agree that every choice has antecedents and all agree, indeed, that this fact puts freedom of choice in doubt. How can a choice be made freely if it is the product of independent antecedents? The response that they make is that some possible antecedents are better than others from the point of view of free choice and that a choice is free to the extent that its antecedents, or at least its relevant antecedents, satisfy the inherently vague condition of leaving it up to the agent.

ANALYSIS 56.4, October 1996, pp. 191–201.

Formulaic compatibilism assumes one or another substantive form, depending on how the two crucial placeholders in this account are filled in. Ask a compatibilist what exactly they think the relevant antecedents are, and ask what exactly they think it is for its antecedents to leave a choice up to an agent, and you will learn all you need to know about the substantive position they defend. Formulaic compatibilism becomes a substantive compatibilism to the extent that these two vague points in the account are made precise.

We try in the sections following to give a sense of the different forms that compatibilism may take by looking at the different precisifications of these notions. There are three main ways, it seems to us, in which the notion of an action being up to an agent may be articulated and these give us three families of compatibilist doctrine. Each of those modes of articulating that notion, however, may be associated in turn with different specific doctrines, depending on what antecedents are cast as relevant. We look at the three families of compatibilist doctrine in the three sections following, illustrating each with some specific forms that it may take. In the fifth section we show that not only does each approach take different forms, depending on which antecedents are seen as relevant; there is an important divide in the ways in which relevant antecedents may be identified and this marks a difference, pertinent to each approach, between a more robust and a less robust variety of compatibilism.

2. *Freedom as underdetermination*

The three ways of interpreting what it is for a choice to be up to an agent are associated, respectively, with the notions of freedom as underdetermination, as ownership and as responsibility. The first interpretation takes a choice to be up to an agent to the extent that the agent could have done otherwise: the future at the time of choice was open or underdetermined. The second takes a choice to be up to an agent to the extent that it is not due to anyone or anything other than the agent themselves; it is a choice that the agent owns, a choice with which the agent identifies, and not something forced upon them. The third takes a choice to be up to an agent to the extent that the agent can reasonably be held responsible for it, being subject to praise or blame for having performed it; the choice is not the product of any independent excusing factor.

The notion of freedom as underdetermination springs from the idea that if someone does something freely then it must be the case that they could have done otherwise: at the time of acting, the future was somehow open. Let it be granted that the things people do have antecedents: lawlike, and no doubt causal, antecedents. If it is going to be possible for an agent who

φs to have done otherwise, then it must be that these antecedents are consistent with the agent's not φ-ing. But how is that going to be possible in the sort of mechanistic universe which the compatibilist takes for granted?

One way in which it may be taken to be possible is by postulating that while the universe is mechanistic, while it conforms to patterns of iron law, those laws are fundamentally probabilistic rather than deterministic. In particular, they are probabilistic in such a way that for anything an agent freely does, the antecedents of that choice do not necessitate it; consistently with those antecedents being just as they are, a different choice might have eventuated. But compatibilists generally allow for freedom in a world that is deterministic and not merely mechanistic. So how can they proceed under this assumption? Taken as a whole, the antecedents of any choice will necessitate that choice under a deterministic picture and compatibilists of this stripe must take the relevant antecedents to be a subset of the totality. But which subset?

The longest established approach, associated with Thomas Hobbes (*Leviathan*, Ch. 21) identifies the relevant antecedents by class as the sorts of hindrances that other people may put in the way of a person's choice. Hobbes's line is to say that a choice of φ-ing is free just to the extent that these sorts of antecedents, these efforts on the part of other people, are consistent with the agent's not φ-ing. To be free in φ-ing is to be let alone in φ-ing: to φ, simply as a result of whatever antecedents in your own history or make-up lead you to act.

But the Hobbesian line is not the only one possible. Imagine that in different conversations about agents and choices – in different perspectives on those matters – different antecedents are taken into account. In ordinary, day-to-day discussions, for example, only familiar psychological and social antecedents are considered whereas in psychiatric discussions those antecedents expand to include childhood experiences and complexes, family relationships, social conditioning, and the like, and in neurophysical discussions they give way to the finer-grained antecedents that we find as we look into the brains and nervous systems of the subject in question. This observation allows us to see how someone may conceive of freedom as underdetermination and identify the relevant antecedents that leave a choice up to an agent by reference to a favoured perspective. The line will be that an agent is free to the extent that the antecedents that can or have to be countenanced in that perspective leave the choice underdetermined.

Suppose that the perspective favoured, for example, is that of everyday conversation about beliefs and motives and obligations and the like. In that case the compatibilist will say that a choice is free to the extent that the

antecedents visible in that perspective underdetermine it.[1] No matter that the antecedents visible in a psychiatric or neurophysical perspective leave no alternatives open. What it is to be free is to be underdetermined by antecedents associated with the other stance. To be free, if you like, is to be free relative to that stance.

3. *Freedom as ownership*

The notion of freedom as underdetermination, just on its own, is not likely to be very attractive for compatibilists. The fact of underdetermination, however understood, is purely negative in significance; it registers an absence of determination without requiring anything positive. The trouble with such a negative way of understanding free will is that it seems incapable of distinguishing between the action that represents a purely random event – inherently random or random relative to certain antecedents – and the action that issues from what we spontaneously imagine as the will or mind of the agent. This problem may be the explanation for the greater popularity among contemporary compatibilists of the conception of freedom as ownership.

The ownership line takes a choice to be up to an agent to the extent that it is not due to anyone or anything other than the agent themselves; it is a choice that the agent owns, a choice with which the agent identifies, and not something forced upon them. Suppose that the relevant antecedents in the adjudication of free will are taken to be the sorts of factors that are salient in the so-called intentional perspective; specifically, beliefs and desires. The ownership version of compatibilism will hold, then, that an agent φs freely just in case their beliefs and desires combine to lead – or at least lead in 'the right way' (see Davidson 1963) – to their φ-ing. The idea will be that the action is free to the extent that it is not brought about in the manner of a reflex, in the manner of an accident, or by the coercion of body or will: it reflects the agent's characteristic way of seeing and valuing the world.

Those who take this version of the ownership approach must face the objection that an agent will not necessarily own or identify with an action just because they act on the basis of their beliefs or desires. What if the

[1] We do not take on the philosophical task of elucidating the defensible notions of perspective that may be brought to bear here. We note, though, that one might think of a perspective as requiring us to attend to certain facts, permitting us to attend to others; but that equally one might think of a perspective as permitting us to attend to certain facts and precluding our attending to others. In the latter sense a perspective precludes a God's eye view. Analogies: The game of blind man's buff cannot be played by an omniscient being. Nor can the game of trust: I cannot rely on the goodwill of someone if I focus on the fact that his neuronal circuitry will determine that he act in a certain way (see Pettit 1995).

desires, or even indeed the beliefs (see Elster 1982), are wanton, in a phrase introduced by Harry Frankfurt (1971)? What if they come and go without invitation or welcome on the part of the agent, so that the agent cannot conceive of the actions in which they issue as characteristically his or hers? This objection readily suggests amendments to the bare belief-and-desire story, such as the amendment favoured by Frankfurt himself. This would require, not just that the action issue in the right way from the agent's beliefs and desires, but that it issue from desires which the agent desires to have and be moved by: that it issue from first-order desires that the agent endorses at a second order. At the very least, the requirement is that the action issues from desires that the agent has some measure of second order control over.[2]

The conception of freedom as ownership scores over the conception of it as underdetermination in representing free choice as requiring positive credentials, not just negative ones. But how does it do in saving the intuition behind the first conception, that with any free choice the agent could always have done otherwise? One well-known approach[3] (though by no means the only one available) will be to interpret what this means, in a more or less deflationary fashion, on the following lines. An agent who acts out of their beliefs and desires in φ-ing could have done otherwise to the extent that they would have done otherwise had they wished: that is, had their desires been other than they were. This reading of the underdetermination point will not satisfy some, on the grounds that it may not have been possible for the agent to have wished to do otherwise.[4] But it is not our place to assess – certainly this is not the place to assess – only to taxonomize.

4. *Freedom as responsibility*

The last broad strategy which compatibilists may try to explore starts from a different angle again. If we consider someone to have acted freely in performing a certain action, then we take the person to be responsible for what they did: to be rightly deserving of praise or blame, for example, depending on whether they did well or ill. Anyone who takes one of the other approaches will struggle to explain why this should be so. But some-

[2] This sort of strategy is also pursued by Daniel Dennett 1973 and 1984. See Vihvelin 1994 for a discussion of this sort of approach.

[3] See Moore 1912, ch. 6, and the papers by Ayer, Aune and Lehrer in Watson 1982.

[4] See Chisholm 1964, p. 27, and van Inwagen's discussion of this style of compatibilist analysis in his 1983. It is not only incompatibilists who will object to this analysis, however. Those compatibilists who take the requirement of second-order control seriously will also, clearly, reject the analysis as too simple-minded.

one who takes the third approach will see it as the most central and basic feature of all. In particular, they will see it as a feature that is inadequately explained under the other approaches, for neither the agent who is given a certain slack by the relevant causal order, nor the agent who acts on the basis of how relevant factors are organized within them, looks just on that account to be answerable to adjudication.

It is common wisdom that 'ought' implies 'can': that is, that if a subject is answerable or responsible to adjudication, then they have the capacity to act as is normatively prescribed. The first two approaches offer accounts in normatively independent terms of what it is for a person to satisfy the 'can' – of what it is to be free – and must try, on that basis, to show that because of satisfying it, people are suitable addressees for the corresponding 'ought'. The third approach reverses this order. It takes the fact of responsibility, the fact of normative adjudicability, as given and it identifies as free will the capacity in human beings that this answerability implies (see Pettit and Smith 1996).

Suppose that we see the antecedents by reference to which we should judge of responsibility as those antecedents that count as good excuses for doing whatever is done: as factors such that if they operated in determining the choice, then the choice was not free. If such excuses are taken to be an intuitively salient class, then it will be possible to say that an agent is responsible, and the agent's choice is free, to the extent that none of those excuses are available: to the extent that those antecedents leave the choice as a matter for which the agent is answerable.

But the excuses approach represents only one of a number of ways of implementing the conception of freedom as responsibility. We may also implement it by arguing, more abstractly, that we have to hold people responsible in law and morality – no society could work without the acknowledgement of responsibility – and that a person does something freely when the antecedents of the action are consistent with their being held answerable in that way for it.[5] The antecedents relevant in the determination of free will under this approach will be those that are visible to the eye of the law or the eye of morality; they will be identified as those to which the practice of law or morality requires us to direct our attention. This line may be combined with a sort of relativism, since law and morality are evolving systems, at least under positivistic conceptions. And as there is variation in the antecedents which law and morality bring into view, so there may also be a variation in ascriptions of free will. At one time, for example, no appeal to certain psychiatric complexes may be allowed to

[5] This argument might build on points in Hart 1949, though it will have to avoid the problems raised by Geach 1960, and acknowledged as problems in Hart's Preface to his 1968.

undermine the ascriptions of responsibility whereas at another time such an appeal may become commonplace.

Yet another form of the responsibility approach, associated with P. F. Strawson (see his 1962), would privilege the participant or reactive standpoint that we spontaneously assume in dealing with others, and not anything as circumscribed as law or morality. When we feel gratitude or resentment, or any of the so-called reactive attitudes, we hold people responsible for what they did: gratitude or resentment would make no sense otherwise. Holding people responsible in this way is an essential aspect of maintaining ordinary interpersonal relations with them. It is suspended only when we come to think of them, in an objective rather than reactive manner, as a suitable case for investigation and treatment: as someone who needs psychiatric care or medication. The Strawson approach is to say that a person does something freely to the extent that the relevant antecedents of the action – the antecedents visible in the participant stance – allow us to maintain a reactive attitude to them; they do not require us to think of the person as a suitable case for psychiatric treatment.

This conception of freedom as responsibility is exemplified, finally, in theories which start from the fact that human beings are essentially concerned with valuation and with tracking the values they countenance (see Watson 1975 and Wolf 1990). The idea here is that as we look on people as valuers we will see them as free and responsible just so far as we can regard them as systematically responsive to those values: just so far as we are not forced to cast them as the playthings of whatever desires happen to strike. Free will under this approach is nothing more or less than the capacity to track values. It is 'orthonomy': the capacity to be ruled by the 'orthos', by that which is right (see Pettit and Smith 1990).

How does the responsibility approach fare in sustaining the intuitions behind the other approaches: the intuition, first, that free choice is underdetermined choice and, second, that free choice is choice with which the agent is identified, choice which the agent owns? To hold an agent responsible in certain choices is to think that it is not inevitable either that they get things right or that they get them wrong – either that they do well or that they do ill – and so it is to believe that there is a sense in which they could have done otherwise: that they do otherwise under contingencies that cannot be discounted as outlandish or unlikely. Again, to hold an agent responsible for certain choices is to think that the path they take is not dictated from outside, by forces with which they are not identified, and so it is also to believe that there is a sense in which the choices are their own.

5. More robust and less robust compatibilisms

The specific versions of each construal of freedom identify the antecedents relevant to free choice in one of two ways. Either directly, as in the approaches which instruct us to look for the hindrances that other people may offer to the agent, for beliefs and desires at the source of the agent's response, or for factors that can be quoted by the agent as excuses. Or indirectly, as in the approaches which focus our thoughts on those antecedents, whatever they are, that are taken into account in ordinary discussions of free will, in the intentional perspective on human behaviour, in legal or moral discussions of responsibility, in the participant viewpoint, and so on. The difference between these ways of identifying antecedents will not be significant to the extent that we treat them as alternative ways of picking out a fixed set of antecedents. To be free is to be free relative to this or that set of antecedents, so it will then appear, and there is no great difference made by whether the antecedents are picked out in a direct or an indirect manner. But we need to signal that there is a different way in which the indirect identification of antecedents may be implemented and that this is of the greatest significance for the nature of the compatibilism defended.

Imagine that we are looking at the ways in which the conception of freedom as underdetermination may be articulated and that we are investigating the possibilities of taking relevant antecedents to be those highlighted in ordinary discussion, or in psychiatric discussion or in neuroscientific discussion. The choice of an indirect strategy will not be of particular significance, if we assume that whichever perspective is privileged, it will remain privileged as we ourselves adopt a different perspective. We will make this sort of assumption if, taking the ordinary perspective to be privileged, for example, we assume that it remains privileged as we ourselves look at psychiatric or neuroscientific antecedents: if we assume that even in psychiatric or neuroscientific contexts, the question of free will remains the question as to whether the ordinary antecedents – the ordinary antecedents, not the psychiatric or neuroscientific ones – leave the agent sufficiently underdetermined.

But it should be clear that consistently with identifying relevant antecedents by perspective, we need not make an assumption of contextual independence. On the contrary, we may think that in ordinary discussion, the ordinary antecedents are relevant, in psychiatric discussion, the psychiatric antecedents, and in neuroscientific discussion, the neuroscientific antecedents. We may take the view that the question of whether someone enjoys free will amounts to a different question, depending on which perspective is contextually indicated. Relative to the ordinary context, perhaps the person is free: perhaps the words 'the agent acted freely' express a truth, as uttered in that context. But relative to another context,

the person may not be free; the words 'the agent acted freely' may not express a truth when they are uttered in that context.

This contextualist variety of compatibilism would make free choice elusive or non-robust in the way in which David Lewis (1993) has recently argued that knowledge is elusive. Lewis maintains that the claim to know that *p* is the claim to have evidence that rules out *not p* in relevant scenarios; but that attending to a scenario is sufficient for its being relevant and so that the claim to know something cannot be maintained in philosophical discussions where sceptical scenarios are introduced. Talk of knowledge is fine in everyday contexts, then, but only because we take so few scenarios into account; talk of knowledge loses a connection with fact as we become more tolerant in allowing how things might just happen to be.

By analogy to the Lewis line, the contextualist variety of the underdetermination approach would save talk of free choice in everyday contexts but only because we take so few antecedents into account in those contexts. Talk of free choice will become less and less likely to be sound as we go philosophical and take more and more antecedents – say, the antecedents visible from the neuroscientific perspective – into account. Lewis's approach to knowledge gives a lot away to scepticism, for it allows that the concept of knowledge becomes inapplicable in contexts where we do not restrict ourselves to considering just some scenarios. And this approach would give a lot away to incompatibilism, for it allows that the concept of free choice would become inapplicable in contexts where we do not limit ourselves to the consideration of just some of the antecedents of the choice. Take up the philosophical stance and, assuming that the universe is deterministic, free will will cease to be anywhere visible.[6]

As the contextualist twist may be introduced to give us a more elusive form of compatibilism than that which is generally associated with the conception of freedom as underdetermination, so it may be employed to give us elusive versions of the other conceptions too. The observation is important. It shows that each of the three main families of compatibilism allows of one form under which it is only the letter of free will that is saved, not the spirit. Contextualist compatibilisms all allow that in some contexts the words 'the agent acted freely' can express a truth. But they allow

[6] Lewis's analysis of knowledge runs on the lines: S knows that *p* iff S's evidence eliminates every possibility in which not-*p* – Psst! – except for those possibilities that we're properly ignoring. The idea is that insofar as one attends to a scenario, one is not ignoring it and, a fortiori, one is not properly ignoring it. Here is an analysis of freedom that deploys a similar strategy: S freely performs an action, *A*, just in case S's doing *A* was not brought about by factors beyond S's control – Psst! – apart from those factors that we are properly ignoring.

equally that that image of a free will fades out as we move to other, usually other more comprehensive contexts, and that free will only retains the reality that goes with an acceptable *façon de parler.*

6. *Conclusion*

When we are led to predicate freedom or free will of an agent, we presumably do so on the basis of believing, explicitly or implicitly, that they satisfy the constraints that we intuitively associate with being free. We believe that someone acts freely, acts of their own free will, only so far as it is true that they could have done otherwise; only so far as the action they perform is one that they identify with; and only so far as the action is one for which they can reasonably be held responsible. The freedom of an action, as we target it in common discourse, is nothing more or less than the property that makes those propositions come out true. We triangulate on that property, as it were, from the standpoints of the three different connections.

The different strategies surveyed in this paper each assign a different hierarchy to the sorts of propositions given and, in the course of doing so, offer different interpretations of them. The first family of approaches makes the proposition about underdetermination basic and, if it does not reject the other two as of no importance, tries to derive them from it. The second gives a similar axiomatic status to the proposition about ownership and tries to derive the propositions about underdetermination and responsibility as theorems. And the third approach attempts a similar ploy, taking the proposition about responsibility as the fundamental axiom of free will.

There are different criteria that are relevant to assessing the different strategies. First, we will want to know which way of axiomatizing the relevant propositions – the free will commonplaces – offers us the best interpretation of those principles, fits best with other general principles that we take as basic and answers best to our intuitions about when particular individuals are free, when not; we will need to know which way of construing free will best satisfies the test, in John Rawls's phrase, of reflective equilibrium. Second, we will want to see how far the different ways of understanding free will direct us to a property that we can really expect to find in actions that are performed in a mechanical, even deterministic universe. And third we will wish the compatibilist to offer us a compelling diagnosis as to why so many philosophers find incompatibilism so compelling.

But the task of assessing the different varieties of compatibilism for their performance on these criteria is not part of our brief in this paper. Our aim

has been to make the various possible sorts of compatibilism visible, not to provide an assessment of their merits.[7]

Syracuse University
Syracuse, NY 13244-1170, USA
jphawtho@syr.edu

RSSS, Australian National University
Canberra ACT 0200, Australia
pnp@coombs.anu.edu.au

References

Chisholm, R. 1964. Human freedom and the self. Reprinted in Watson 1982.
Davidson, D. 1963. Actions, reasons and causes. *Journal of Philosophy*, 60: 685–700.
Dennett, D. 1973. Mechanism and responsibility. In *Essays on Freedom of Action*, ed. T. Honderich. London: Routledge. Reprinted in Watson 1982.
Dennett, D. 1984. *Elbow Room*. Oxford: Clarendon Press.
Elster, J. 1982. *Sour Grapes*. Cambridge: Cambridge University Press.
Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20. Reprinted in Watson 1982.
Geach, P. T. 1960. Ascriptivism. *Philosophical Review* 69: 221–25.
Hart, H. L. A. 1949. The ascription of responsibility and rights. *Proceedings of the Aristotelian Society* 49: 171–94.
Hart, H. L. A. 1968. *Punishment and Responsibility*. Oxford: Oxford University Press.
Lewis, D. 1993. Elusive knowledge. University of Melbourne Philosophy Preprints, No. 26/93.
Moore, G. E. 1912. *Ethics*. London: Williams & Norgate.
Pettit, P. 1995. The cunning of trust. *Philosophy and Public Affairs* 24: 202–25.
Pettit, P. and M. Smith. 1990. Backgrounding desire. *Philosophical Review* 99: 565–92.
Pettit, P. and M. Smith. 1996. Freedom in thought and action. *Journal of Philosophy*, 93: 429–49.
Strawson, P. F. 1962. Freedom and resentment. *Proceedings of the British Academy* 68: 1–25. Reprinted in Watson 1982.
Van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.
Vihvelin, K. 1994. Stop me before I kill again. *Philosophical Studies* 75: 115–48.
Watson, G. 1975. Free agency. *Journal of Philosophy*, 72: 205–20. Reprinted in Watson 1982.
Watson, G., ed. 1982. *Free Will*. Oxford: Oxford University Press.
Wolf, S. 1990. *Freedom within Reason*. New York: Oxford University Press.

[7] Our thanks to Michael Smith for discussion of the points made.