Article:
Steward, H (2014) Replies to Randolph Clarke, John Bishop and Helen Beebee. Res Philosophica, 91 (3). 547 - 557. ISSN 2168-9105

10.11612/resphil.2014.91.3.13

Replies to Randolph Clarke, John Bishop and Helen Beebee

(i) Randolph Clarke

Randolph focuses in his response on the main argument offered for Agency Incompatibilism and suggests that it suffers from a serious shortcoming. In fact, I think he identifies more than one potential shortcoming in the material he discusses; and I shall concede that some of those he identifies do indeed exist. But I shall also do my best to defend my work against what I think is the nub of the most significant of Randolph's objections, namely, that some of the important notions which are put to work in my argument - notions like 'up-to-usness' and 'settling' - are susceptible of different interpretations, and that nothing has been done to suggest that the interpretations I need to get the argument for Agency Incompatibilism off the ground are ones which truly pertain to our concept of agency.

Randolph gives, I think, a very good summary of the position for which I wish to argue. It is indeed not the concept of substance causation *per* se, but rather the concept of *agency*, that I regard as being in problematic tension with the doctrine of determinism. Substances can certainly cause things in deterministic worlds; to use my terminology, there could be *movers* in such worlds, and nothing I say suggests, or was intended to suggest otherwise. It is agency that is the problem. I look in detail at the idea of substance causation only in the service of an attempt to show that there is nothing essentially incoherent about the very idea that a substance can be the cause of something, not in order to make incompatibilistic mileage out of the very idea of substance causation. I alleged that it was part and parcel of our concept of agency that agents are settlers of at least some matters at the times at which they act, in a sense of 'settle' according to which a matter cannot come to be settled at time *t* if it has ever been settled at a time prior to *t*. And Randolph grants that the phrase 'settle whether' can be used to express the idea that I am here attempting to characterise. But he also notes – as indeed I do myself in the book – that we can also talk about settling in contexts in which it seems clear that no indeterminism is being presupposed. The same arguably applies to a whole host of other concepts – 'up-to-usness', 'power to refrain', 'open alternative' – to which incompatibilists might wish to appeal when trying to say what it is they believe. Language is horribly slippery here; almost every word for which one reaches in the attempt to articulate the conception of agency I wish to allege we possess can be given a deflationary, compatibilist reading. And given that this is so, it is evidently not enough just to say that agency involves settling and hope that that alone will do all the argumentational work. Randolph's challenge, essentially, is the question: even if we accept that the idea of agency is connected somehow to the idea of settling, what is the argument for thinking that this is so on the *incompatibilist* reading of 'settling' that I attempt to characterise, rather than a compatibilist one? Randolph suggests that I offer no such argument. That would indeed be a serious shortcoming. But Randolph's claim is not correct.

Before I say why it is not correct, I want to make two concessions to Randolph. The first is that he is right to suggest that a move I made in attempting to differentiate what I called the weaker (compatibilist) from the stronger (incompatibilist) conception of settling, by invoking *time,* will not work, and for the reason he gives – that one can settle at a given time (in one sense) something that was already settled at a prior time (in another sense). All the ambiguities in the concept of settling can simply be recapitulated for the concept of settling a matter *at a time*, so that is clearly a move that at the very least needs considerable refinement if it is going to do any work in isolating the distinctive incompatibilist sense of 'settle'. The second concession, which I fear I have found myself having to make rather frequently to compatibilist opponents of my position since the publication of

A Metaphysics, is that it is quite true that I am much too quick to suppose that all compatibilists are likely to reach for an understanding of settling that is dependent for its details on structures provided by the Causal Theory of Action. Many compatibilists have pointed out that there is a much richer range of options available, and they are right about that. So I regret my over-exclusive focus on the Causal Theory, although I should like to say in my defence that it still does seem to me true that its presuppositions continue to exert a very powerful influence on many compatibilistic ways of thinking about agency.

Now for the central point on which I think Randolph is mistaken. Randolph supposes, I think, that the whole content of my answer to the question what justifies the claim that our concept of agency has the character it does is supposed to come in Chapter 3. He notes that in that chapter, having raised the suggestion that there might be compatibilistic ways of understanding the central notion of settling, I move directly to focus very heavily on one particular *version* of this compatibilist conception – that offered by the Causal Theory of Action; and as I've said, he's obviously right that that is too limited a response to function as an entirely general answer to all compatibilists whatever. But it is as though Randolph supposes that the question what justifies the assumption that the concept of agency really does have the incompatibilist lineaments I allege it does is supposed to be entirely answered in Chapter 3. That, however, is not the case. The question what is the structure of our concept of agency, and whether it does it in fact resemble the concept I attempted to delineate in Chapter 3, is an *empirical* question, and it is in Chapter *4* that I intended to take on the burden of answering it, by appeal to evidence from developmental psychology, as well as more traditionally philosophical appeals to intuition, conceptual relations, etc.

Here is what I say at the beginning of Chapter 4 (pp.71-2) to introduce the argument of that chapter:

"I shall suggest, in what follows, that the normal development of the infantile processing of animal activity results in the eventual emergence of a mature conception of agency that has roughly the following features:

(i) An agent can move the whole, or at least some parts, of something we are inclined to think of as *its* body;
(ii) An agent is a centre of some form of subjectivity;
(iii) An agent is something to which at least some rudimentary types of intentional state (e.g. trying, wanting, perceiving) may be properly attributed;
(iv) An agent is a settler of matters concerning certain of the movements of its own body in roughly the sense described in Chapter 2 – i.e. the actions by means of which those movements are effected cannot be regarded merely as the inevitable consequence of what has gone before.

Feature (iv), of course, will turn out to be particularly crucial for me, since it is in virtue of (iv) that the *agency* concept can be seen to embody a *prima facie* commitment to indeterminism. But it is also the inclusion of (iv), I anticipate, that is likely to prove most controversial. In view of this, I shall devote quite a large portion of the chapter to the consideration of what may be said in justification of its inclusion".

This extended quotation, I hope, makes it clear that one of the aims of Chapter 4 was precisely the defence of the claim that Randolph claims I leave entirely undefended, namely, the assumption that our concept of agency does indeed have the incompatibilist structure I allege it does. And Randolph does not say anything at all about the material in Chapter 4.

Now, it may be, of course, that Randolph says nothing about the material in Chapter 4 because he finds it unpersuasive. Perhaps it *is* unpersuasive – but offering unpersuasive arguments is not the same as failing to realise that any need to be offered at all. Moreover, while I fully accept that nothing said in Chapter 4 can be regarded as a knock-down argument, I repeat the point that what has to be established here is something *empirical* – namely, the actual structure of the deep concept of agency which permeates the organisation of human thought. And where empirical questions are concerned, particularly ones the experimental investigation of which is at a relatively early stage, it might be foolish to suppose that knock-down arguments are going to be forthcoming. What I took myself to be doing in Chapter 4 was amassing a body of evidence, partly from the developmental psychology literature, partly from intuitions often expressed in the philosophical literature, which I took, at any rate, to be strongly *suggestive* of the view that our concept of agency has the features I suggest it does. I concede that there is more to be said, and I anticipate that more experimental evidence of the sort, for example, that Shaun Nichols (2004) has attempted to provide (and which I actually found wanting in Chapter 4) might be forthcoming in the fairly near future. But the provision of a knock-down argument was not really what I was trying to provide. My main aim was to show something that I think modern ways of setting up the free will debate has obscured – namely that the concept of *agency*, thought of as a widely shared animal power, might itself be a powerful source of incompatibilist intuitions and arguments, and that without so much as even mentioning ideas such as freedom and moral responsibility, one can get incompatibilist worries off the ground. I do of course also believe that Agency Incompatibilism is true, but because it incorporates a claim about the nature of a concept whose lineaments are contested, its defence cannot proceed entirely *a priori*. I am pretty convinced that my concept of agency involves the incompatibilistically construed power of settling, and because I think the concept is deep, and because I think it is a cognitive given, I believe yours does too, even if, at the reflective level, you consider yourself a compatibilist. But we will not settle that issue – if readers will pardon the pun – merely by trading intuitions. We will settle it only by adopting the methods of developmental psychology and looking to the findings of cognitive science, which was what I aimed to do, to the best of my abilities, in Chapter 4.

(ii) *John Bishop*

John's insightful response kicks off with some general reflections on the relative merits of my own approach, which seeks to invoke a variety of agent causation, and his own preferred version of the view he calls 'realisationism', which invokes the Causal Theory of Action (CTA). He then proceeds to examine in more detail my precise criticisms of his defence of the CTA. I will try to say something, in what follows, about both aspects.

John notes that in a sense, I am no less of a realisationist than he is. I think this is true and important, and John's response gives me a good opportunity to affirm it. Agent causation, according to me, really is realised by the activity of the parts of my body; my doing things is in a certain sense constituted by neurons firing, electrical potentials changing, and all the rest of it. Agent causation is not a strange input from some other ethereal realm into the physical world; agent causation is simply the influence of a whole animal on its own parts. And since animals are physical, their influence on things is also physical, and conducted, therefore, entirely by means of changes – and indeed non-changes - taking place in its bodily parts. And this is also something John believes. So what is the difference between us? Why do I call myself an agent causationist and a libertarian, while John calls himself a causal theorist of action and a compatibilist?

One issue on which, I think, we are in considerable disagreement concerns the general plausibility of the notion of top-down causation. John believes that on what he calls a 'strict reading', it is hard to accept that there could be any such thing. By a 'strict reading', he means an account according to which, "when it acts, the agent controls its relevant parts (such as neurons and hormones)" (p?). Now, I agree that it might be odd to *say* that agents (under normal circumstances) control such things as neurons and hormones, but it is the same oddness, in my view, that would attach to saying of a driver who knows nothing about what is under the bonnet (hood!) that s/he was controlling the clutch plates separating the engine from the transmission, when the left-hand foot pedal is pressed. It is odd, and perhaps misleading, to speak of someone controlling something of which they have no direct awareness, since we usually specify the objects of control by means of descriptions which highlight the features which are consciously controlled by the controller. But it is unclear that it is *false* that a driver in fact controls the clutch plates (by means of controlling the clutch pedal), when s/he drives. And even if we decide that it *is* false – that the semantics of the word 'control' forbids its use under such circumstances – it still seems true that the agent does indeed *cause* movement to occur in the clutch plates when she pushes the clutch pedal, whether s/he knows it or not. And it is, recall, the notion of top-down *causation,* not the notion of top-down control which we are supposed to be considering here. It is true that the phenomenology of action does not generally give many clues to the underlying mechanisms by means of which we bring about motion in our own bodies; but neither does the phenomenology of driving give clues to the mechanisms we engage when we press the pedals, turn the steering wheel, and so on. But that should not stand in the way of the truth of claims to the effect that we do indeed cause certain (unknown) events to happen when we drive.

It is of course true that our relation to our own bodies is in many ways not at all like our relation to such things as the cars we may drive. For the actions by means of which we cause things to occur inside our own bodies must also in some sense or other be found to occur in those bodies, by anyone who is a realisationist. This is what makes the problem so difficult philosophically. How are the mere caused motions to be properly distinguished from the causing actions? How is it that we can be justified in finding the one kind of event to be merely the effects of our activity, and the other the genuine manifestation of ourselves in the world, when both kinds of event can be viewed, at the neural level, as merely the movings and changes of neural parts, in response to prior movings and changes in other such parts? Well – that is the question I tried to answer – admittedly programmatically - in Chapter 8, invoking a range of resources in the attempt to make the prospect of developing a workable account of top-down causation seem less hopeless than I think it is generally apt to seem to most realisationists, and in my darker moments, than it is apt to seem even to me. I stressed the importance of the fact that the animal's input not be seen as something *prior* to whatever neural processes initiate and then monitor and control the relevant bodily movement or change. The important relations between animal and parts have to be synchronous, not diachronic. Another important element of the account is the idea of *coincidence.* Some of the most impressive features of our world are due to the fact that certain extremely complex combinations and orderings of lower-level entities - in the case of animal agency, for example, things like neurons firing in different regions of the brain - come into being *together* at just the right time to ensure that certain other complex combinations and orderings are facilitated at the next stage. How is this to be arranged for by the world? Just citing *prior* complex arrangements also existing at the lower level, together with the laws that govern them, leaves mysterious and unresolved the vast coincidence that this perfect conjunction of circumstances represents. We seem to need to raise our eyes to higher levels of reality to arrive at the possibility of resolving this coincidence – the coincidence occurred, we need to be able to say, because e.g. the animal was trying to dance! – and so of course her thus trying brought the relevant parts of her motor system into line, not subsequently to the

trying, as it were, but as part of what it *is* for an animal to try. I also tried to try to loosen the grip of the thought that lower level sufficient conditions at every stage must dictate the next, and the idea that supervenience of higher upon lower levels of reality alone seems consistent with the thought that the evolution of reality *over time* might be due to the way things are at the higher, rather than the way things are at the lower levels of reality. Thus, though there is a sense in which the activity of an agent is entirely realised in the activities of her parts, there will be no understanding the evolution of the properties of these parts from moment to moment without invoking the activities of the agent. I think John might actually agree with me about quite a lot of this. But the difference between us is that I do not see how the agent's input could truly count as such, as something genuinely stemming from the agent herself in the special way that action demands, if determinism were true. If determinism were true, then the agent's input would seem simply to stem from whatever are the prior events which necessitate her trying, and those from events prior to that, and so on. Whereas what seems required, to me, is that the top-down form of influence required for a naturalistic account of agency be an indeterministic power, such that it remains an open possibility at each moment whether or not it will be exercised.

What about the second part of John's commentary? – the material on the deviant causal chains objection to the CTA? John is, I think, right to point out that Hillel-Ruben's report of John's conditions on the causal relations intrinsic to agency is inaccurate, and right, too, that this makes a difference to the question whether Hillel-Ruben offers a true counterexample to John's theory. But though I regret not noticing the inaccuracy explicitly, and must apologise to John for propagating Hillel-Ruben's original mistake, I think I did *in effect* consider the question whether John's account might survive Hillel-Ruben's attack by considering its virtues as an account merely of *sufficient* causal conditions for agency (as John agreed, in responding to a comment in discussion made by Michael Bratman, would be quite enough for a viable CTA). That is to say, I do in effect consider the question whether, *provided* John's 'feedback to the central processes of M condition' is met, (and irrespective of whether or not feedback *also* goes to the central processes of a second agent), we might be assured of having a case of agency.

But my answer to this question was 'no' – this is not assured – and I still don't really see why what I said in the book about this is not right. John's feedback condition was originally introduced as a means of ruling out cases in which omniscient and benevolent interveners assure the connection between M's basic intention to do a and behaviour b, and ensure that the mechanism is suitably sensitive, but intuitively without M having actually to *do* anything. But if we are to rule out such cases, John's (ii), read as it must be read to avoid falling foul of the Hillel-Rubin case, will be insufficient. For suppose feedback from behaviour b *does* reach M's central processes, but that what then happens to it is simply that it is read off by the intervener, who then ensures that behaviour b is properly adjusted so as to continue to expedite the fulfilment of M's intention. I see no reason to suppose that such causation might not be 'sustained', nor indeed why things that counted as 'servosystems' might not be involved (though perhaps there is a reason that I have not discerned?). And so if there was a worry about omniscient and benevolent interveners to begin with, then surely there will be a worry here too. What seems required is not that the feedback merely reach M's central processes but that it then be utilised *by M* to produce the wanted behaviour. But then we have failed to avoid having a clause in the account which does not essentially advert to the need for the causation involved to be causation *by M*, and hence to presuppose what we are attempting to analyse.

(iii) *Helen Beebee*

Helen's commentary is that wonderful thing – a discussion of one's work which makes one see more clearly the structure of one's own thinking, and the assumptions underlying one's position. The centrepiece of Helen's response is that one important assumption I make (and which she thinks requires more justification than I give it) is that *lawful sufficiency* requires *causal completeness* – and that that assumption is not obligatory. In what follows, then, I shall try to say something about this central claim.

In a way, I think I am very much in accord with Helen's inclination to believe that the solution to certain conundrums about causation, necessitation, sufficiency, overdetermination, etc., is likely to lie in careful scrutiny of these connected concepts, and in the disruption of very natural but nevertheless non-obligatory assumptions about the way in which they relate to one another. And in some ways, I think her suggestions about where the crucial moves might be made are not a million miles away from my own. I suggest, for example, in the last chapter of *A Metaphysics*, that we should challenge the idea that "an effect of a given kind is always fully accounted for, metaphysically speaking, one proximal causally sufficient conditions for its occurrence have been provided" (p.236). In formulating this suggestion, I use the notion of 'causally sufficient conditions' rather than that of 'lawfully sufficient conditions', and the notion of something's being 'fully accounted for, metaphysically speaking', rather than the notion of something's being 'completely caused' – but mutatis mutandis, my thought, I think, is actually quite closely related to Helen's suggestion that lawful sufficiency does not require causal completeness.

But the linguistic differences between the terminology for which each of us reaches may be significant. In using the idea of conditions which are *causally* sufficient, I mean to record my thought that if such conditions were ever to obtain, they would be related to their consequent effects by a relation of causal necessitation, in the sense that the latter *could not fail* to follow from the former – and I then seek the space that is required for agency in the thought that it might be, nevertheless, that we need metaphysically to account for the existence of certain lower-level conjunctions of circumstance by raising our eyes to items found only at higher levels, such as whole animals. It is *the animal*, on my view, which causes the right microphysical conditions for action to occur to be present in the first place (though by way of a variety of causation that it is essential to think of as synchronic and top-down, rather than diachronic and reducible to the activity of parts on parts). Helen prefers, I think, to think of the preceding lower-level conditions as related by things we can regard as *laws*, to general facts concerning what then follows from them – but then seeks the space that is required for agency either in the thought either that lawful sufficiency might not amount to causal sufficiency (this is the variety of solution she attributes to List and Menzies) – or else in the thought that causal sufficiency at lower levels might not exclude the existence of proper counterfactual dependencies, of the sort we generally regard as underwriting causal relations, between higher-level facts. It may perhaps be helpful, then, if I say a little bit about my worries about these alternative forms of solution to the problem of causal exclusion, and my reasons for thinking that they do not answer all the questions one might have about the role of the agent.

Both of the varieties of solution to the problem of causal exclusion that Helen favours seem to be premised on a broadly counterfactual conception of causation. I suppose I have always felt that although causation and counterfactuals are evidently very closely related, the idea that causation might just *consist* in the holding of suitably specified counterfactual relationships is implausible – that causation certainly involves the existence of relationships of various sorts from which counterfactuals flow, but cannot itself be reduced in any way to the holding of counterfactuals. And

so perhaps that is one source of concern about the solutions proffered. I am more inclined than is Helen, I think, to be quite a thoroughgoing realist about causation, whereas I think Helen is probably more Humean than I am inclined to be, and therefore likely more sceptical about the idea that power, necessary connection, and like ideas refer to real features of metaphysical reality. Another worry is perceptively identified by Helen herself. Counterfactual relationships, as Helen rightly points out, are first and foremost relationships between the sorts of things I call *mattering* causes – they are specified as relationships amongst *propositions.* And while I concede that the sorts of moves to which Helen adverts might very well suffice to show that certain propositions concerning higher level causes might matter causally to whether a given effect occurs or not, even given lawful, or even causal, sufficiency at the lower level, I am not convinced that these moves will suffice to provide adequate guarantees about the role of the agent. What they will make possible is such things as this: that it will be causally relevant to the fact that the agent went to the shops that she wanted to buy some bread, and thought she'd be able to buy some there. And it is, of course, important to make room for this sort of higher-level causal relevance. But that facts pertaining to the agent and her mental states should be causally relevant to what she then does, does not, by itself, guarantee that the agent *is* an agent, any more than the fact that facts pertaining to a stone's mass (say) can be causally relevant to the fact that it breaks a window, shows that the stone is an agent. And what I am looking for is a solution that guarantees not merely the causal relevance of higher-level facts, but the special kind of causal *efficacy* of the higher-level object which is the agent – a kind of efficacy which does not reduce to the connected efficacies of its various related parts.

It is, moreover, the fact that the efficacy involved in agency is of a special kind which provides the answer to the question Helen raises about why I do not simply rely on my own causal pluralism in order to insist that necessitating lower-level matterers do not compete with causally powerful substances, and hence do not exclude agents from exercising their distinctive powers, any more than they prevent substances such as cars from demolishing walls. The answer to this question is that *mere* causal pluralism here is too flimsy a resource on which to rely to underwrite the special form of substance causation which is agency. Animal agency is simply more demanding, metaphysically speaking, than mere substance causation – and causal pluralism therefore cannot supply the whole of the answer to the question how it is possible. Neither can it do so even in conjunction with an adequate solution to the causal exclusion problem. These resources are both necessary, but not even jointly sufficient. Because of the need for agential settling, the only metaphysical combination that really seems to me capable of giving us what we need for agency is causal pluralism, plus top-down causation, plus indeterminism.

Moreover, in view of Helen's final objection, it needs to be stressed that these last two components of the metaphysical mix, though related in my view, are not related in quite the way Helen suggests. It is not my view that we need to appeal to indeterminism in order *quite generally* to make room for top-down causation, as Helen supposes (when she objects that this is a claim that places very strong demands on the laws of nature). I simply *agree* that we might be able to make perfectly good sense of certain varieties of top-down causation, even within purportedly deterministic scenarios. What I doubt is merely that we will be able to make sense of the special variety of top-down causation that is *agency* without rejecting determinism. I invoked wheels and whirlpools in an attempt to shake faith in certain bottom-up orthodoxies concerning causation and explanation, and to show how supervenience alone need not dictate an entirely unidirectional account of how the determination of facts at one level by facts at another must proceed. I did not mean to imply these cases, too, require indeterminism for their understanding – which is, as Helen herself suggests, a rather implausible claim. It is only in the case of agency where I think an argument for the existence of an

indeterministic form of top-down causation might be forthcoming, based on the notion that an agent must *settle*, as well as cause, what happens.

I would like to end, though, as Helen herself does, on a conciliatory note. As I said earlier, I do not think our views are actually as far apart as all that – and I wonder whether some of our remaining disagreements might lie in our different understandings of what, precisely, determinism itself involves. In the original APA symposium, from which both Helen's comments and this reply have evolved, Helen suggested that perhaps I had been insufficiently clear in *A Metaphysics* about what exactly determinism *is*, and I think she was right about that. Her conception, I think, is probably a bit different from mine, because I think she has a somewhat different understanding of the sort of thing a law of nature must be. At the end of her commentary, Helen says that she has long believed that the issue surrounding the correct account of the laws of nature is of crucial importance to the free will debate. Helen's commentary has made me see that this is indeed a central issue – and that one cannot sensibly proceed, as I had rather hoped to do in the book, without being much more specific than I generally managed to be about the precise content of the thesis of determinism. And she has also made me see that my central commitment is actually to *indeterminism*, rather than to incompatibilism. Many incompatibilists *derive* their indeterminism from their incompatibilism – it is free will, and free will alone, that leads them to deny determinism. Whereas for me, indeterminism (interpreted as the denial of DPD) has its own independent plausibility – and free will is thereby (happily!) much more easily accommodated. Helen has said all this much more clearly than I managed to do myself – and I am very grateful to her for making it plainer to me.

Bibliography

Bishop, John. 1989. *Natural Agency*. Cambridge: Cambridge University Press.

Hillel-Rubin, David. 1991. 'Review of John Bishop: *Natural Agency*'. *Mind* 100: 287-90.

List, C. and Menzies, P. 2009. 'Nonreductive Physicalism and the Limits of the Exclusion Principle'. *Journal of Philosophy* 106: 475-502.

Nichols, Shaun. 2004. 'The folk psychology of free will: fits and starts'. *Mind and Language* 19: 473-502.

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