Published paper:
Steward, H (2012) *The Metaphysical Presuppositions of Moral Responsibility.* The Journal of Ethics, 16 (2). 241 – 271.

http://dx.doi.org/10.1007/s10892-012-9127-5

The Metaphysical Presuppositions of Moral Responsibility

There is a certain irony in the fact that I have been asked to contribute to a special issue of a journal which is devoted to 'Current Work in Moral Responsibility Theory'. The irony is that in so far as I can claim to have worked at all on the topic of moral responsibility, it is primarily to urge that certain questions with which the concept of moral responsibility has become hopelessly entangled, need to be approached afresh, and from a purer, metaphysical perspective. One major concern about moral responsibility has always centred around its compatibility, or otherwise, with the thesis philosophers call 'determinism' – which we can take, for present purposes, to be the thesis that "for any given time, a complete statement of the (temporally genuine or non-relational) facts about that time, together with a complete statement of the laws of nature, entails every truth as to what happens after that time".[1] But why is determinism potentially in tension or contradiction with the idea that we might be morally responsible for some of our actions? The answer usually given to this question is that moral responsibility requires alternative possibilities of a kind that determinism might appear to rule out. For an agent to be morally responsible for doing something, we might think, she needs to have been able to have done something different from what she did do. And if one asks, in turn, why *that* might be so, one might be likely to reply: 'because otherwise, she cannot be properly blamed or punished for what she has done, if what she has done is bad, or wrong; and (more contentiously) she cannot be appropriately praised, if what she has done seems good, or right'.[2] For it would not be *fair*, it might be thought (though it might perfectly well be expeditious, or practically useful, or whatever) to blame someone for something she could not have avoided doing (and perhaps praise would likewise be unmerited in a case in which an agent cannot help doing as she does).[3] Fairness – a moral concept – thus assumes a central place in the debate – and the central question in the area can quickly seem to be about the compatibility of *that* – of fairness in blaming, praising, punishing, etc. – with determinism.

It is not that I do not think it is interesting and important to ask about whether determinism is consistent with the fairness of these practices by means of which we signal personal or societal approval and disapproval of what people do, or fail to do. But to connect moral responsibility with alternate possibilities through essentially moral notions, such as fairness and desert, is to connect it through only one of the possible routes by means of which these two concepts may be associated. What I have tried to urge in my work is that there is another route available – a route that goes not through the idea of fairness, but rather through the idea of *agency*.[4] The connection stems from the fact that it is plausible to think that being an agent in the first place is a necessary (though not, of course, a sufficient) condition for being the kind of entity which could be morally responsible for anything. And though many currently very widespread views both about what actions are, and about what determinism is, make the claim more difficult to defend than it ought to be, it is possible to provide good reasons for supposing that agency itself is inconsistent with determinism. I call this view – that agency itself is incompatible with determinism - *Agency Incompatibilism*, and it is the view I shall be attempting to explain and justify here. Clearly, if (i) determinism is incompatible with agency; and (ii) the

[1] I have borrowed this particular definition from Fischer (2006:5).
[2] See e.g. Wolf (1990) and Nelkin (2008) for the view that praiseworthiness, though not blameworthiness, is consistent with not having been able to do otherwise.
[3] Again, the case of praise is more contentious, and I concentrate henceforth, for simplicity's sake, on blame.
[4] See especially my (2009a) and (2012).

existence of agency is a necessary condition of the existence of moral responsibility; then (iii) determinism will also be incompatible with moral responsibility – so that we have a route to incompatibilism between moral responsibility and determinism that is not the one that goes directly through the idea of fairness. Moreover, I believe, it is the better of the two routes – for it can withstand recent objections to the claim that moral responsibility requires the capacity to do otherwise, which are largely based on the assumption that the first route is what is supposed to provide the argument. I shall not argue here for (ii), taking it to be tremendously plausible, though doubtless not unquestionable. I shall focus on the case for (i), which has not generally had the airing it deserves.

In the first and longest part of the paper, I shall try to explicate and justify Agency Incompatibilism. The most important part of this task is the characterisation of the conception of agency on which it depends; for unless this is understood, the rationale for the position is likely to be missed. I shall accordingly take quite some time to set out what I take to be the orthodox philosophical position concerning what it takes for agency to exist (section (i)), before explaining, in (ii), why, and how, I believe that orthodoxy should be challenged. Then, in sections (iii)-(v), I shall consider the relation between my views and those of others writing on the issues of free will and moral responsibility, in three crucial and inter-connected areas: (1) the question how *animals* should figure in the philosophy of action; (2) the question what the lesson is of 'Frankfurt-style' examples; and (3) the distinction between so-called 'leeway' incompatibilism and 'source' incompatibilism. Sections (vi)-(viii) will then consider and respond to various objections to Agency Incompatibilism, including the claim that to embrace the conception of agency that makes incompatibilism plausible is to beg the question against the compatibilist, and also the worry that determinism is an empirical thesis which ought not to be straightforwardly falsifiable by such *a priori* reasoning as Agency Incompatibilism appears to involve. I will also tackle the worry that Agency Incompatibilism is committed to the existence of an unintelligible and/or naturalistically impossible variety of irreducible agent causation. By the end, I hope it will be clearer why one might want to be an Agency Incompatibilist – and why, in many ways, the position has a claim to be a more coherent and naturalistically respectable version of libertarianism than any of the others which are currently on offer in the literature.

(i) *The concept of agency: the contemporary orthodoxy*

Why might anyone think that agency was incompatible with determinism? Before we can answer that question, we need to get clear what might be meant here by 'agency'. We can make a very slight beginning (and one to which no one is likely to object) by saying that agency is the power to *act*. But what is it to have the power to act? Which sorts of things can possess that power, and under what circumstances are they to be found exercising it? It is here that I believe most modern philosophical accounts provide a quite inadequate answer, which is apt to conceal from view the specialness and distinctiveness of action as a causal category, and with it, the appeal of Agency Incompatibilism.

An answer to the question what it is for something to exercise the power to act that would be fairly typical of the contemporary philosophy of action literature might go as follows. An agent may be said to have acted when, and only when, a bodily movement of hers has been produced in a certain distinctive way.[5] That way involves the bodily

[5] There are at least two sorts of reasons for thinking that this formulation, in attempting , as it does, to offer a necessary, as well as a sufficient condition for the occurrence of an action, might be too strong. But I do not myself think either line of reasoning is compelling. The first is based on the thought that *omissions* ought to

movement's being causally produced by certain special sorts of mental states and events – beliefs and desires, for example, must interact to give rise to intentions, which may go on, in some cases, to produce decisions – which perhaps in the presence of beliefs of the form that *now* is the time to act, in turn produce the requisite bodily movements. Views of this general sort constitute a class of positions which may be regarded as variants of a general idea for which it will be useful to have a name – let us adopt one that has already been pressed into service to do this very job, and call such views variants of the *Causal Theory of Action* (CTA). The condition is usually added by causal theorists that the causal production of movement by mental states must occur 'in the right sort of way' – it having been widely recognised that the problem of deviant causal chains renders the existence of the necessary extensionally conceived causal relations by themselves insufficient to constitute an instance of action.[6] But though to my knowledge, no one has ever given an account of what 'the right kind of way' amounts to that has revealed itself to be both proof against counterexamples and clearly non-circular, few causal theorists have been deterred from the assumption that some such account of action must be correct. The precise details of the causal chain required for the production of action vary from philosopher to philosopher – e.g. some insist on the inclusion of decisions, others think them strictly unnecessary in every case; some allow for representational or informational states less 'full-blown' than beliefs, etc.. For my purposes, though, these differences are much less important than what all views of the sort in question have in common. What they have in common is the idea that it might be possible to analyse the concept of an action in such a way that no terms referring to such things as *agents* appear in the analysis. Instead, the analysis invokes merely a cast of interacting mental states and events. And once this basically event-based view of action is in place, it is obvious why no one would be likely to worry about the consistency of the idea that actions sometimes occur, with the thesis of determinism. The CTA is tailor-made to provide the wanted consistency. Even if the world consists exclusively of chains of interlinked events and states interacting so as to produce the next with all the inevitability that determinism supposes, provided the relevant causal relationships do indeed exist between bodily movements and their causal precursors (and provided the causation happens 'in the right kind of way', whatever that may be), there is no reason to think that *actions* could not occur under determinism. But I want to suggest that it is highly questionable whether this view of the concept of agency is satisfactory. Here is not the place to attempt properly to refute the view; I have tried to do that more comprehensively in my (2012). I will, however, indicate a few of my reasons for being doubtful that it is correct, before setting out the view of agency which I prefer, and explaining why I think it has the better claim to be regarded as the fundamental concept which is at the heart of our thinking in this area.

I have already mentioned one reason to be worried about the CTA – and that is the persistence of the troublesome, unexplicated 'right kind of way' clause in the causal analysis. Its persistence is a sign, I think, that something is deeply wrong with the causal theory – for 'in the right kind of way' sometimes seems, in the end, to be susceptible of no further clarification than can be provided by saying merely that it is the way that ensures that the

---

count as a variety of action, yet are excluded by the 'bodily movement' formulation; but in my view, we do better to agree that omissions, though they are certainly things for which an agent may be responsible, are not themselves exercises of agency (though *decisions* to omit to φ may be), and require separate treatment. The other suggests that *mental* actions cannot be accommodated by the formula. But in fact, it is not clear that this is the case. Movements which occur in the brain are 'bodily', presumably – and I see no *special* reason (that is, a reason over and above the general reasons for objecting to the formula), therefore, why such a thing as a decision or a mental calculation, or whatever, could not be thought of as involving the production of a bodily movement by a special sort of cause, or set of causes.

[6] For discussions of the problem of deviant causal chains in connection with the Causal Theory of Action, see Davidson (1973), Peacocke (1979), Bishop (1989).

agent was the source of the bodily movement that ensued. And unless we can do better than that, we have clearly not succeeded in providing the wanted event-based analysis. Another legitimate concern is that the category of action is made unreasonably narrow by what I have called the 'over-mentalization' in which the causal theory indulges.[7] The basic point here is that not everything one might want to call an action seems to have the sorts of mental antecedents that the causal analysis makes definitive of the occurrence of an action – think, for example, of actions that seem to be done for no particular reason, such as my meandering aimlessly up and down the room (where the requisite beliefs and desires might be hard to light upon – do I have to 'want' to be meandering aimlessly up and down, for instance?), or which are done absent-mindedly, or habitually, or which are 'sub-intentional', like the numerous shufflings, jigglings and twiddlings (some quite rhythmic and controlled) in which most of us engage to a greater or lesser extent throughout the day.[8] Or think of the activities of animals (of whom more later) – where it might in some cases seem far more obvious that the animal is an agent performing actions, than that those actions are produced via states it would be obviously right to call 'beliefs', 'intentions', etc. Which sorts of actions are difficult for the causal theory to accommodate will of course depend on the precise details of the particular version on offer; and of course it is always an option to deny that the relevant bodily movings really *are* actions. But this denial comes at what is potentially the very large cost of excluding huge tranches of the activity in which we (and other animals) engage on a daily basis from the category of action. And it is worth asking the general question why it should be thought so obvious in the first place that the idea of action is the idea of a movement produced by some event or state which is distinctively *mental*. For many actions seem to take place apparently entirely in the absence of such things as forethought, deliberation, conscious attention, and so on – and can be thought of as caused and rationalised by some particular set of beliefs and desires only after a variety of often implausible Procrustean manoeuvrings – the beliefs in question end up having to be 'unconscious' or 'implicit', for instance, and the desires have, in some cases, to be for things the agent clearly does not want in any ordinary sense. To put it bluntly, when one thinks about the extent to which our lives (and the lives of other animals) are dominated by activities which are unthinking, habitual, routinised, absent-minded, unplanned, spontaneous, etc. (or is it just me?), one might be forgiven for wondering why it is so often assumed without much argument at all that specifically *mental* antecedents are the hallmark of an exercise of agency. It is also a reason for suspicion, I think, that the 'movement-with-mental-cause' conception of action looks like a relic of a basically Cartesian way of thinking of humans and their doings as neatly decomposable into mental aspects and physical ones, the mental bits causing the physical when an action occurs. It is true that almost no one believes that the mental aspects are to be thought of as essentially non-physical any longer. But it must surely be a mistake to suppose that we can arrive at a satisfactory philosophy of mind simply by replacing states of the Cartesian soul with states of the physical brain and leaving everything else more or less as it was on the Cartesian view. One of the lessons to be learned from the failures of Cartesianism is that the *conceptual* psycho-physical divide itself is a mistake from which we need to struggle to free ourselves – not merely that we have to be careful to avoid substance dualism and ensure that everything is metaphysically rooted in the motions and changes of matter.

But the *real* problem with the CTA is that it *cheats*, and does not really offer the reductive conceptual analysis which it is supposed to provide. It smuggles in conceptual apparatus which is taken from a theory of agency it purportedly eschews, while trying to

[7] See my (2009b).
[8] See O'Shaughnessy (1980) for the concept of a sub-intentional action.

pretend that it invokes nothing more than a range of supposedly innocuous event-causal notions. Take, for instance, that notion of a 'bodily movement'. That concept incorporates already the rich idea of a *body*; it is only of some of the things we meet with in experience that we say that they 'have' bodies, and hence are capable of making 'bodily movements'. No clock or computer or mobile phone 'has' a body – we only introduce the conceptual scheme of owner, on the one hand, and body owned, on the other, in certain special kinds of case – cases, roughly speaking, in which we detect *animacy*. The concept of a specifically *bodily* movement, then, is already dependent upon the implicit understanding that we all share, of the distinction between things with bodies and things which are not suited to be the possessors of bodies. But that means that the concept of a bodily movement is not in fact the innocent event-causal notion it purports to be. For it is not clear that we can characterise the notion of a movement which is, specifically, *of a body*, rather than of any old kind of thing, without relying on this distinction between the body-possessors and the variety of other things which can move. The distinction between things which 'have' bodies and things which do not, indeed, has a good claim to be simply *identical* with the distinction between agents and non-agents – to think of a thing as an agent arguably is just to think of it as something which bears the ownership relation to a body, something which stands to that body and its bodily parts in the relation of controller and director. In utilising the notion of a bodily movement, therefore, the CTA is helped to arrive at the right answer to the question in what circumstances an action may be said to have occurred – it is indeed true, on my view, that it is only where *bodily* movements are found that genuine *actions* may be said to have been performed by things which are true *agents*. But it arrives at this admittedly correct answer only by helping itself illicitly to a notion – the notion of a body - which already presupposes that of an agent in offering its purportedly agent-free analysis of agency. The same, I would argue, is true also of the various mentalistic concepts by means of which the causal theory tries to delineate the characteristic *antecedents* of action – beliefs, desires, intentions and the rest. These are concepts for the everyday understanding of which we rely heavily on the assumption that there is a believer to do the believing, a desirer to do the desiring, etc. – an agent, in other words, to whom these states belong, and to whom they are attributed on the basis of patterns of activity. It cannot be straightforwardly assumed that we have any idea what we are talking about in utilising these concepts within a context in which they are supposed actually to *do duty* for the believer, the desirer, etc. – and hence where no separate subject of these various states can be straightforwardly presupposed. The slippery concept of a 'state' is a culprit here, for states can be states 'of' things, so that there is nothing evidently peculiar about the idea that beliefs, desires, etc., fall into the category; but they are also often thought and talked of in philosophy as though they could potentially be quite independent, causally efficacious entities, understandable without essential reliance on the idea of the entities of which they *are* the states. But even if both conceptions of statehood are legitimate (and I have my doubts about the second[9]) we must be careful to see that we do not slide between them without noticing, and thereby create an illusion of comprehensibility out of what is in reality a morass of confusions. The fact that beliefs, desires, etc., are 'states' in the first sense does not imply that we properly understand the idea that they could hang around in the world on their own, so to speak, without a subject to *have* or be in them.

For all these reasons, then, I am very doubtful about the adequacy of the CTA to provide what I think it has pretensions to provide – viz., a purely event-causal way of thinking about what has to happen in order that an *action* should have occurred. But what is the alternative? Is there another way of thinking about what it is for an agent to act without running the risk of tangling oneself up with noumenal selves and supernatural unmoved

[9] See my (1997).

movers? I think there is – and in what follows, I shall try to explain what I think is a far preferable approach to the concepts of action and agency. My view is not exactly a straight *rival* to the CTA – for I simply do not accept that it is possible to do what the CTA supposes it is possible to do – and that is to give a *conceptual analysis* of action in other (and specifically, in event-causal) terms. I shall not, then, be attempting to provide an alternative set of necessary and sufficient conditions for the occurrence of an action – for it is no more obvious that there *is* any such set of necessary and sufficient conditions, than in the many other cases in which conceptual analyses of important concepts have been fruitlessly attempted (e.g. knowledge, perception, meaning). Rather, I start from the idea that the concepts of an agent and of an action are best approached as concepts in what one might think of as the *psychologist's* sense.

(ii) *The concept of agency: an alternative view*

Psychologists working in the fields of developmental and cognitive psychology tend to operate with a conception of concepts according to which *conceptual* identity is more or less independent of the identity of *words*.[10] To see what is at issue here, consider the word 'animal'. Young children are often strongly predisposed to deny, when asked, that human beings are animals – they have learned to associate that word 'animal' with an extension that does not include human beings and have to be taught that the word 'animal' can also be used to denote a more inclusive concept of animality, roughly corresponding to the idea of the animal *kingdom* within which humans ought clearly to be included. But does this mean that prior to this teaching they lack this more inclusive animal concept? Not at all. For its possession is discernible, despite the adamant denial of most of these children that humans count as animals, in such things as their patterns of inductive projection. Animals and humans alike are attributed such properties as the capacity to feel pain, to breathe, to eat, to move themselves around the world, thus indicating that these children are possessors of some version of the animal kingdom concept already, long before they understand that the word 'animal' can be used to denote it. Concepts, on this view of what they are, are *deep*. They represent ways in which things tend to be classified together for the purposes of reasoning and thinking by the human mind – and there may or may not be individual words in individual human languages which map onto them precisely.

When I speak of agency, I mean to be speaking of one of these deep concepts. It seems to me quite clear – and empirical research in developmental psychology confirms it[11] - that human beings are predisposed from a very early age to regard some of the things they meet with in experience in a way extraordinarily different from the way in which they regard certain others. These special things are regarded as (i) sources of their own motion; (ii) centres of subjectivity (albeit, in the case of simpler creatures, of a rudimentary kind); (iii) targets for the application of a raft of special 'mentalistic' concepts, of which some of the more basic are perceiving, wanting and trying to get, but which also include more sophisticated notions such as believing and intending; (iv) possessors and controllers of things we call 'their' bodies; (v) things which are potentially suitable referents for personal pronouns like 'he' and 'she'. For ease of reference, I shall call the assumptions encoded by (i)-(v) 'the agency scheme'. None of the apparatus embodied in these assumptions is routinely and automatically invoked in the cognitive comprehension of inanimate entities - though of course there are many special cases even here, mostly involving only a partial

[10] See e.g. Carey (1985).
[11] See e.g. Premack (1990), Gelman (1990), Leslie (1994), Gelman, Durgin and Kaufman (1995), Spelke, Phillips and Woodward (1995), Gergely and Csibra (2003).

application of the scheme, and for which there are special explanations (e.g. children's relations to their dolls, teddies, etc.; beliefs inculcated through religious or cultural teachings, such as the belief in some cultures that trees have 'spirits', etc.; the tendency to interpret certain movements of shapes on a computer screen as e.g. a big circle 'trying' to contact a smaller one, etc.). But these interesting and special cases notwithstanding, there can be, I think, no doubt whatever that living things above a certain level of complexity are accorded a very special sort of treatment by our cognitive apparatus, a treatment which is only accorded to inanimate things in exceedingly partial or avowedly metaphorical ways. And it is the categorisation imposed by this cognitive apparatus to which I believe we need to look in order to characterise the concept of agency – for it is here in the deep and murky operations of cognitive systems we have no choice but to deploy, and whose functioning is already empirically discernible in infants at the tender age of three months, that the main lineaments of this extraordinarily important concept are to be discovered.

That such an important categorisation is indeed imposed by our cognitive systems has, of course, in a sense been quite widely recognised in philosophy, as well as in psychology, for some time. There has been a huge flurry of interest over the past thirty years or so, for example, in the idea that we are predisposed to understand the causal workings of other human beings by reference to what is often called 'folk psychology' – that is to say, by way of the special raft of mentalistic concepts mentioned in (iii) above. But the philosophical treatment of folk psychology has, I think, been problematic. Philosophers have tended to represent folk psychology as though it were simply a theory about an interacting causal network of events and states – beliefs, desires, decisions, etc. - which is imposed by our cognitive systems on certain favoured portions of reality. But in my view, this characterisation of folk psychology is untrue to the scheme we actually operate, and wholly underestimates the distinctiveness of the structure of concepts we bring to bear on the universe in the relevant favoured places. The scheme we actually operate, in my view, involves regarding *the agent herself* as the source of her own movements - no doubt movements which are made, in many important and central cases, because she has certain purposes and believes that these movements will serve in some way to further these ends. But the desires and beliefs that are implicitly ascribed by this assumption of purposiveness are conceived of as dependent phenomena. They do not hang about in the world, all by themselves, according to the agency scheme; they have to be *possessed* by a subject who then acts on the basis of them. And moreover, it is a crucial part of the agency scheme that reasons, however, good, desires, however strong, intentions, however resolute, decisions, however firm, are by themselves all entirely insufficient to get an agent into voluntarily produced motion. All the intending and deciding in the world is not enough to get a movement made. It is part and parcel of the agency scheme that if any voluntary motion is actually going to happen, the agent also has to *act,* which is a further move. It is part and parcel of the agency scheme that action is never an inevitable consequence of any given prior constellation of mental states and/or events. Action involves *execution* – something the agent has to effect at the time of action and which can never be produced inevitably merely by an antecedent set of prior conditions. The appropriate action therefore may or may not follow on from predisposing, antecedently existing desires, resolutions, decisions, or whatever, depending on what the agent actually does. This is another way of saying that the agent is conceived of by the agency scheme as a possessor of what is sometimes called *two-way* power – the power to φ *or not to φ*. Exactly what will occur is not settled in advance by antecedent states and events, according to the agency scheme. It is settled by the agent at the time of action by means of an exercise of a two-way power. Actions are thought of as what I call *settlings,* by the agent, and at the time of action, of what will occur in respect of certain

portions of the agent's body. It is the action itself, and not anything prior, which settles the details of precisely what will happen.

What is the evidence that this is our scheme? Most of it, I think, is staring us in the face. It is possible to view the long history of the free will problem itself as evidence that we are unable easily to marry our understanding of the type of causality implicit in the agency scheme with the broadly Humean models of production of one thing by another that we have come to regard as sufficient for our dealings with the inanimate world. One might also mention the ready sense we are able to make of the idea of weakness of will (as folk, though not always as philosophers). We seem to have no difficulty with the idea that no matter how good one's reasons for doing something, and however strong one's motivations, actually getting it done is another matter, something which may or may not follow on from antecedently predisposing leanings, desires, etc., depending on how the agent actually acts. And there is some limited experimental evidence too. Shaun Nichols, for example, has investigated the question whether folk psychology contains any commitment to what he calls 'agent causation' – and his results suggest an affirmative answer.[12] Admittedly, agent causation is a tricky concept and not everyone agrees about how best to characterise it; certainly, there are problems with Nichols's own characterisation, which makes it not entirely straightforward to interpret his results in terms which I regard as satisfactory.[13] But still, his evidence strongly suggests that agents are thought of by us as entities with the two-way power to do a certain thing *or not*, in contrast to inanimate entities, which are thought of rather as normally constrained absolutely by the circumstances in which they find themselves to do precisely whatever it is they in fact do. Further empirical work is doubtless needed – but there is enough evidence of various sorts already around to make it a very plausible hypothesis that the agency scheme has the character I have alleged it does, whereby an agent with discretion to act or not to act is presupposed.

Now, it might be, of course, that this agency scheme is confused or incomprehensible or inconsistent with what science tells us about the workings of biological organisms, including ourselves. Whether it is or not is a question I shall come on to tackle shortly. But it does only harm if our philosophical preconceptions about what does and does not make sense, or what can and cannot be squared with modern science, is allowed to interfere from the beginning with our descriptive account of what the folk psychological scheme actually *is*. That is, I think, what has happened in philosophical discussions of folk psychology. It has been taken for granted by most of those trying to describe the explanatory scheme we bring to bear on the understanding of animate beings that of course the idea of an agent, an *entity* with the causal power to bring about movement in a way that does not simply reduce to those movements having been brought about by states of the entity, events occurring in the entity, etc., does not make sense. The agency scheme has then been characterised in such a way that it is not guilty of what is taken to be this metaphysical confusion. But the task of describing our cognitive predispositions is one thing, the task of assessing them for coherence another. Of course, it might seem unlikely, for all sorts of reasons, that we should have an incoherent conceptual scheme – and certainly it is a good methodological principle that the attribution of

---

[12] See Nichols (2004).

[13] In particular, one of the beliefs which counts for Nichols as constitutive of belief in agent causation is the belief that actions are caused by agents. But in my view, this way of putting things is problematic. For it leads to the question *how* agents cause their actions – where the answer must not be 'by acting', since that would set us off on an infinite regress. But if this is not the answer, it is not clear what the answer could be – we seem left with no alternative but to embrace the dubiously comprehensible idea that agents may somehow cause actions without doing anything at all. This predicament is best avoided entirely by insisting that actions are *not* caused by agents – that they are rather the causings, by agents, of *other* sorts of event (such as, for example, in the first instance, movements and changes in bodies). The answer to the question how agents cause those other events then is indeed 'by acting'.

incoherence to the structure of a cognitive framework should be avoided if there is any reasonable alternative. But it is a methodological constraint that ought not to assume a role so important as to be permitted entirely to distort our account of how it is we categorise the world. I believe that our 'folk' ways of thinking about action and activity have been quite grossly misrepresented by philosophers seeking not to attribute to the folk a view that they take to involve confused metaphysics. But, first, it is not obvious that the folk view might *not* involve what, on reflection, might turn out to be confused metaphysics – and therefore we should not allow our philosophical preconceptions to interfere with the description of the folk scheme. And second, the metaphysics in question might in any case turn out to be less confused that these philosophers have supposed.

I hope it is already obvious why, once the concept of agency is characterised in the way I have suggested, there is an issue about the compatibility of agency with determinism. If an agent is genuinely to possess at the time of action the power to do something *or not*, and thereby settle with that action what is to occur, it would seem that what will happen cannot *already* have been settled in advance by the past and the laws. It is very natural, therefore, to think that there must be more than one physically possible future if an agent is to have it in her power either to φ or not to φ at the time of action. If agency of the sort that the agency scheme is committed to exists, it requires at the very least some rather fancy compatibilist footwork to avoid the conclusion that the world cannot be deterministic if there is agency in it. The most natural conclusion to draw is that indeterminism is a necessary condition of agency – and hence, in turn, a condition of moral responsibility.

That, in barest of bare outline is the case for Agency Incompatibilism. The outline is of course *only* bare, at present – and much remains to be said both about how the general picture is to be filled in and about how various pressing objections are to be met. In order to make some progress now with the first of these tasks, I want to turn in the next section to highlight some of the differences and similarities between my views and those of others who have written recently on the topic of moral responsibility. Then, in sections (v)-(viii), I shall try to tackle what I regard as the most worrisome and obvious objections to my position.

(iii) *Animals and Humans*

One of the interesting differences between the way I approach the issue of free will and the way most others approach it, is that on my view, it is not only humans whose capacities and characteristics present difficulties for deterministic visions of the universe. Animals, too (at any rate ones which exceed a certain crucial level of complexity) are treated in the relevant special way by the cognitive systems which I allege are the root of the concept of agency – they are regarded by those systems, I maintain, as sources of their own movement; they are thought of as centres of subjectivity (they can be in pain, for instance, and can see and hear things); they are things to which at least some serious subset of the full range of mentalistic concepts can be applied (e.g. they can know certain things, they can want an object and try to get it; they can act with purpose); they are regarded as possessors and controllers of things we call 'their' bodies; and they are regarded as appropriate targets of personal pronouns, such as 'he' and 'she' – at any rate, when we have sufficient information to make the choice of gender possible. Their actions, moreover, are treated by our cognitive systems as exercises of two-way powers – as settlings, by the animal, at the time of action of what will then happen to its own body. Determinism is thus difficult to square not only with the morally responsible varieties of agency we take ourselves to possess – but also with the much simpler forms of animal power which are found far more widely distributed across the biological realm.

I do not deny, of course, that cultural factors and education of various sorts may come to impinge hugely on our confidence in the applicability of the agency scheme to non-human creatures; and so that we can easily come to deny, on reflection, that a wasp or a starling or even an orang utan truly merits its special attentions. And I do not deny either that such reflection is important and appropriate – and that we *ought* to subject our native predispositions to scientific and philosophical scrutiny. But – to repeat a point already made above - we should not allow the effects of the scrutiny to obscure the profile of the agency scheme itself. I believe it is a scheme we are inclined to ascribe almost as readily to many animals as we are to apply it to ourselves – they, like us, are naturally thought of as beings with goals, desires and a subjective point of view on the world. If we later become Cartesians, or behaviourists, or cautious cognitive ethologists, or simply people influenced in one way or another by one of more of these intellectual traditions, we might come to view the agency scheme – and perhaps especially its application to non-human animals - with much scepticism. But viewing it with scepticism is one thing, refusing to recognise its existence entirely is another. Moreover, I would suggest, our *default* presumption ought to be that the agency scheme locks onto differential workings that really are present in nature. We have the scheme; a sensible hypothesis must be that this way of thinking about animals has evolved because it is useful. And the simplest and most straightforward – though admittedly, not the only possible - explanation of why it is useful is that the distinction between the workings of the animate world and the workings of the inanimate realm which it imposes on our thought answers to certain important features of reality – that the workings of the animate world really *are* different in interesting ways from the workings of the inanimate. The default presumption should be, then, in my view, that there really are such things as agents and that they do indeed have the two-way powers that they agency scheme attributes to them. If we are to be persuaded that there is, after all, no foundation to the agency scheme, or that the scheme itself is conceptually confused, metaphysically flawed, or simply inconsistent with what science tells us about the nature of the universe, that will require arguments. I shall consider such an argument in section (viii) and will suggest that it is not anywhere near as powerful as it has sometimes been taken to be.

The idea that animals possess two-way powers might admittedly be very easily dismissed if it could be shown to conflict with the obvious truth that animal life (including human life) takes forms which are hugely constrained by the operations of animal instinct. But understood as I believe it should be understood, it does not do so, That animals have two-way powers does not imply that they are free not to bother trying to escape from predators they have spotted, not to chase easy prey, not to drink when thirsty, eat when hungry, and so on. It is important to be clear about the level of description at which the two-way power that Agency Incompatibilism insists upon operates. On my view, it is perfectly possible for there to be φ-ings which are actions, such that the agent could not have avoided φ-ing, and which are even such that the agent could not have avoided φ-ing *then* (at the very time at which she in fact φ-s). And instinct is likely to be one very important type of factor which leads to the occurrence of such unavoidable act-types – it might, for example, be impossible for a bored cat to refrain from stalking a mouse it has just spotted in some nearby grass. But, I maintain, it will still be possible for the cat to execute the stalking in any one of a variety of possible ways – by stalking along route x or route y, a little more or less slowly, stopping to reassess the situation more or less often, etc., to hunker down at location A or location B, etc. Even when acting on the promptings of raw instinct, the cat retains powers over the precise organisation and ordering, in the service of her ends, of the movements and changes in her body which go to constitute her activity. The alternative possibilities one needs for action are ones deriving from the necessity, if one is to be an agent, of having power over one's body – of having the power in respect of at least some of the particular movings of limbs, digits, etc.,

or other changes in one's bodily state, which go to constitute one's φ-ing, not to have made those very movements or changes. And this implies that even if one is in circumstances such that one cannot refrain from making *some* movements and changes of the sort that go to constitute an action of type φ (because, for example, of some instinctual necessity) one always has the power not to make the very ones one in fact makes at the very times at which one makes them, provided one is genuinely acting in the first place. The truth about the relation between agency and alternative possibility is therefore not the simple one that in order for a given φ-ing to count as one's action, one has to have been able not to φ. It is the more complex one that in order for a given φ-ing to count as one's action, the φ-ing in question has to have *some* description as a V-ing, say, (e.g. as a moving by S of S's body in precise manner M) such that the agent was able not to V. And the existence of instinct does not preclude animals from having two-way powers of *this* sort.

To say that non-human animals have two-way powers is not, of course, to say that any non-human animals are morally responsible agents. Possession of the power of agency is only a necessary and by no means a sufficient condition of moral responsibility; and therefore, so far as moral responsibility is concerned, I agree wholeheartedly with the more or less universally held position that human beings are the only animals that have it. Nevertheless, if one is an Agency Incompatibilist, the issues surrounding moral responsibility take on a somewhat different shape and structure from the one they are normally taken to have. Many other authors writing on moral responsibility, for instance, begin their discussions by reflecting on the contrast between morally responsible agents and things which are not held to account for their activities. Haji, for example, begins *Moral Appraisability* with this distinction:

> we are willing to adopt, and do in fact adopt, a whole cluster of attitudes toward persons – such as moral abhorrence and resentment, moral admiration and forgiveness, *some* of the so-called "reactive attitudes" – which it seems inappropriate to have toward creatures like koala bears, which we don't take to be morally responsible agents ... I might have good grounds to believe that Kate's kitten, Kitty, was causally responsible – she played a causal role – in the untimely death of Golda, the neighbour's pet goldfish, but it would be inappropriate, it appears, to blame Kitty or to punish her for her deed in a way in which we would blame a cruel youth for draining Golda's bowl.[14]

But this starting point has led to a tendency to concentrate the discussion of moral responsibility around the conditions which set us apart from other creatures – for of course, if we may be accorded moral responsibility for our actions, while non-human animals may not, it is interesting to ask why that is so. What is it about us, one might wonder, which makes us appropriate targets of praise and blame, when other animals are not?

Haji, considering this question, notes that two sorts of factors are frequently thought to undermine moral appraisability – factors which have to do with the agent's *ignorance* of what it is s/he is doing, and factors which have to do with something like *force* - for instance, if an agent is compelled to do something by another agent, or by what is often called an 'irresistible desire'. Though Haji, I think, is mainly concerned with factors which can undermine an individual's moral appraisability on a particular *occasion*, the same two-fold categorisation of factors might seem relevant to the question what sorts of features preclude agents from being subject to moral appraisability more generally. Non-human animals, for instance, it might be suggested, are precluded from bearing moral responsibility both by lacks

[14] Haji (1998: 3).

relating to intellect and understanding, and by lacks relating to a certain sort of incapacity. Only we humans, it might plausibly be suggested, truly understand moral concepts, the difference between right and wrong, etc., and so only we can be expected to have the knowledge necessary for being held to account for our actions. But considerations of force might also be thought pertinent. Only human animals, it might be supposed – and indeed perhaps only a certain privileged subset of them – are able to transcend the dictates and promptings of such things as desires and instincts, and are thus able to avoid acting under a certain kind of compulsion. A picture is thus in danger of emerging (given this starting point) in which animals (and perhaps also young children, and other human beings supposed to be lacking in what is vaguely called 'reason') are treated as subject to a psychological variety of determinism which we might then be apt to imagine that only rational human beings could transcend.

Agency Incompatibilism opposes this picture. It insists that we and at least the higher animals belong on the *same* side of a divide (or, better, perhaps, at the same end of a spectrum) which separates us from merely mechanistic systems. The Agency Incompatibilist can accept that non-human animals are indeed doubtless *more thoroughly* constrained by the operations of instinctive drives than we are, and thus are, in a sense, less free. But the differences here are matters of degree – and neither they nor we operate entirely deterministically. Alternate possibilities are always available to higher animals, in virtue of the discretion distinctive of agency – discretion which allows (at the very least) for such things as the taking of different routes through space to a target destination; different timings and orderings of activities; different means of achieving a given end. Animals are precluded from responsibility not because they, unlike us, are deterministic systems, but because they lack the understanding that would be required – and perhaps (we can concede) the *degree* of freedom from instinctual demands that would be necessary. But crucially, alternate possibilities are not entirely absent from their lives. To say that would be to deny the applicability to them of the agency scheme, to deny that they ever a*ct*.

This different perspective makes Agency Incompatibilism far more appealing, in my view, than many traditional varieties of incompatibilism. For anyone with broadly naturalistic inclinations, and respect for evolutionary and biological science, a view according to which human beings are the only macroscopic objects allowed to escape an otherwise entirely all-encompassing deterministic net can seem deeply unattractive and smacks of suspicious special pleading for our own species. In such a context as this, compatibilism must surely be the position to which anyone at all naturalistically inclined is bound to default. But Agency Incompatibilism offers the possibility of locating human beings squarely within a resolutely evolutionary perspective, while preserving the valuable incompatibilist conviction that our activities are quite unlike the activities of such things as vacuum cleaners, lawn mowers or even computers, and cannot be understood entirely in terms of the same deterministic explanatory schemes as will suffice for mechanistic entities. Moreover, it becomes possible to locate the real source of what is truly most puzzling about the free will problem – the existence of entities that things can be *up to*, things which are more than mere locations for the deterministic interactions of various events and states - in the comfortingly deep resources provided by developments in biological evolution, rather than the comparatively flimsy structures of discursive rationality. It is not easy, admittedly, even given these deep resources, to explain what needs to be explained here – how there can be any entity such that the influence it has over its own parts does not merely reduce to the influence of *parts* on parts – I shall say something more in section (viii) about what sorts of metaphysical resources seem to be needed in order to make sense of the very idea of an agent, as I have characterised the concept. But it seems to me much more promising to suppose that many of the basic developments which have made creatures with two-way powers possible are developments

already found much lower in the scale of evolutionary complexity than human beings, than that they are the products merely of intellectual capacities such as those relating to reflection, deliberation and the conscious discernment of reasons.

(iv) *Frankfurt-style Cases*

A very great deal of recent discussion concerning moral responsibility and determinism has focused around so-called 'Frankfurt-style' cases.[15] In a Frankfurt-style case, an agent is described who appears to have acted in such a way that we feel inclined to hold him accountable for what he has done. And yet circumstances are such that we are tempted to say that the agent in question could not have done otherwise. Imagine, for example, that Gunnar has conceived an intense dislike for Ridley and plans to shoot him.[16] Cosser, who also has reasons for wanting Ridley out of the way, is pleased to hear of Gunnar's plan, but is worried that Gunnar will not carry it through to completion. Being an excellent neurosurgeon, he is able to implant a device in Gunnar's brain which he, Cosser, will activate if there is any sign of Gunnar's resolve beginning to wane. Activation of the device will cause in Gunnar an 'irresistible desire' to carry out the shooting. But as it happens, there is no need for the intervention. Gunnar goes ahead and shoots Ridley in any case, and Cosser never has to do anything at all.

Under these circumstances, our intuitions tend to suggest, Gunnar is just as responsible for his action as he would be in any entirely ordinary case. Since no *actual* interference by Cosser occurs, he is morally responsible for what he has done. And yet, we might be inclined to think, Gunnar could not have done otherwise. For had he wavered at all in his resolve, Cosser would have intervened and caused him to shoot Ridley after all. Gunnar could not, therefore, have done other than shoot Ridley.

Frankfurt-style cases such as this are intended as counterexamples to what has come to be known as the Principle of Alternate Possibilities:

(PAP) An agent is morally responsible for what she has done only if she could have done otherwise.

But if (PAP) is in fact false, as these examples appear to show, the main argument for supposing that moral responsibility must be inconsistent with determinism seems to fall by the wayside. If agents can be morally responsible *despite* being unable to do otherwise, perhaps we do not need to worry about determinism any longer - or perhaps, at least, we do not need to worry about it so far as its consequences for moral responsibility are concerned. Maybe (some authors have conceded) there might still be worries about our freedom, or about whether we could truly be the *source* of our actions under determinism – so perhaps determinism would still be something we had reason to hope was not true. But there would, at any rate, be no particular reason to feel that determinism threatened moral responsibility, and that might be a very welcome result.

This is the conclusion that has been reached by John Fischer[17] – and it has been widely endorsed by many others writing on the topic of moral responsibility. Fischer defends a position he calls 'semi-compatibilism', according to which determinism is compatible with moral responsibility, although it is incompatible with the ability to have done otherwise. In what follows, I want to explain what the Agency Incompatibilist has to say about Frankfurt-

[15] Following the first presentation of such a case in Frankfurt (1969).
[16] This example is taken from Van Inwagen (1983), 162-3.
[17] See e.g. his (1994) and (2006); and also Fischer and Ravizza (1998).

style examples, and why they do not support the conclusion that Fischer (and others) have taken them to support.

I think it must be accepted that in the imagined case, Gunnar could not have done other *than shoot Ridley*. But there is no reason to accept that he could not have done other *than perform an action of shooting Ridley*. 'Shoot' is a verb under which events that are not actions at all can fall – I can shoot you, for example, by accidentally dropping a gun which fires and unfortunately discharges a bullet into your head. In this case, I have shot you, although there has been no *action* of shooting on my part – there is nothing of which I have been the agent, no chain of activity of which I have been controller and director. And therefore even if it is correct to describe what happens in the possible world in which Cosser intervenes as a world in which Gunnar has shot Ridley, it does not follow that it is a possible world in which Gunnar has performed any action.[18] We might introduce a special verb 'shoot_A' , which is stipulated to have application only in cases in which a shooting which is an action has occurred. Then we might say that the relevant alternate possibility which remains open to Gunnar is that he could have done other than shoot_A Ridley. He could have refrained from shooting_A Ridley, simply by not doing so within whatever time-frame Cosser is disposed to allow, since what would then have occurred would not have constituted a shooting_A on Gunnar's part at all.

Fischer might respond that provided Cosser's intervention is made in the right sort of way, there is no reason to deny that what happens in the alternative scenario constitutes an action on the part of Gunnar. For Cosser is supposed to operate not by working Gunnar's limbs like a puppeteer, but rather by bringing about the sorts of mental states in Gunnar which are definitive of action – states like desires and beliefs, for example – which might then go on to produce the shooting. But of course this response already presupposes the Causal Theory of Action, which the Agency Incompatibilist will not want to accept. She will insist that unless the shooting can be seen as an exercise of some kind of two-way power on the part of Gunnar, it cannot be an action of Gunnar's. If it is an action of anyone's it is an action of Cosser's – since it is Cosser who has the power to settle whether the shooting will or will not occur. For some chain of events to constitute the activity of an agent, it has to be under that agent's control and direction. But the chain of events which would have occurred had Cosser intervened is initiated not by Gunnar, but by Cosser – and presumably will have also to be kept on track by Cosser if it is to be truly clear that Gunnar cannot e.g *stop* the chain of causation at some crucial point. Given these facts of the case, it is implausible that Gunnar counts as the agent of an action in this case.

Unfortunately, though, the principle that underwrites the position that it seems possible and natural for the Agency Incompatibilist to adopt in this particular case is vulnerable in turn, I think, to further sorts of counterexample. That principle might be stated as follows: for an agent to be morally responsible for φ_A-ing, it has to be the case that she have been able to refrain from φ_A-ing. But there are cases in which it seems very plausible to suppose that an agent has acted in a certain way – and yet in which she could not have done other than act in that way. Examples may be provided by cases of what Frankfurt has called 'volitional necessity'.[19] Suppose, for example, that my children are trapped in a burning house. Might it not be plausible to suppose that I am simply incapable of refraining from running in to attempt to rescue them? – at least if it is obvious to me that there is no chance whatever of their being rescued unless I do? But it seems most implausible that I do not act when I go in to attempt to save them – and if I succeed, it is also fairly plausible (though there is room for argument about it, no doubt) that I am deserving of praise for having done

[18] See Alvarez (2009) for a similar argument concerning Frankfurt-style examples.
[19] Frankfurt (1982: 86).

so. I think it is not possible, therefore, to hold onto the position that in order for an agent to be morally responsible for φ_A-ing, it has to be the case that she have been able to refrain from φ_A-ing. But it does seem possible to hold onto a more moderate position.

We have already met the more moderate position, indeed, in meeting the objection that animals do not have alternate possibilities because they are in the grip of instinctual necessities. In meeting that objection, I suggested that the truth about the relation between agency and alternative possibility is not the simple one that in order for a given φ-ing to count as one's action, one has to have been able not to φ. It is the more complex one that in order for a given φ-ing to count as one's action, the φ-ing in question has to have *some* description as a V-ing, say, (e.g. as a moving by S of S's body in precise manner M) such that the agent was able not to V. And we can now appeal to this formulation also to deal with cases of volitional necessity and other troublesome cases in which we feel inclined to say that there is a sense in which the agent could not have done otherwise. We may simply concede the point – there *is* a sense in which the agent could not have done otherwise. But there is also a sense in which what occurs remains unsettled until the agent acts. A bit of terminology will prove useful here to formulate the claim I want to make. Let us say that an event or state of affairs whose occurrence or obtaining at a given time *t* is necessitated by certain events and states of affairs prior to *t* together with the laws of nature is 'historically inevitable'. The claim I want to make is that it is impossible that an event that was historically inevitable could be an action.

Note that this does not imply that *facts about what we will do* cannot be historically inevitable, given our motivations and circumstances. What I have said implies that there are indeed sometimes facts about what we will do which are historically inevitable (though these facts will be much thinner on the ground than compatibilists generally tend to suppose). It might, for example, be historically inevitable that I shall attempt run into the burning house at *t*; or that some cat will attempt to catch a mouse that has caught its eye within some given limited time-frame; or (to take another kind of case) that a heroin addict needing a fix badly and presented with a needle, will go ahead and inject the drug before a certain amount of time has elapsed. But I insist that many alternative possibilities will remain available to the agents in question in all these cases, provided that what occurs is indeed an action on the part of the agent. I might, for example, try to enter the burning house via the front door or the back, searching first upstairs or first downstairs, calling as I go, or not, using these words or those, etc.; the cat may wait a shorter or a longer time before pouncing, take this route or that route around an obstacle, etc.; the heroin addict may inject into his right arm or his left, etc. What is crucial to action, on my view, is that any instance of it constitutes an exercise of the two-way power of an animal to settle which movements and changes will occur in the parts of its body over which it has any sort of control. Where everything of this sort is already settled antecedently, the agent makes no contribution, and so what occurs cannot be an instance of action.

I think at this point it will be very natural for those used to thinking about the alternative possibilities requirement in a very different way, to insist that even if the existence of such alternate possibilities as these is conceded, they cannot be of any importance. To use a term coined by Fischer, they might not seem to be 'robust'. It has become a very widely accepted principle that it is insufficient for the defender of PAP merely to show that an agent in a given case indeed has certain alternate possibilities – it is essential that she also show how those alternate possibilities *ground* the agent's responsibility in the case in question. Here is Fischer making the point in connection with Van Inwagen's suggestion that the agent in a Frankfurt situation could at least have avoided bringing about the individual event (or 'consequence-particular') he does in fact bring about:

… my basic worry is that this alternative possibility is not sufficiently *robust* to ground the relevant attributions of moral responsibility. Put in other words, even if the possible event at the terminus of the alternative sequence … is indeed an alternative possibility, it is highly implausible to suppose that it is *in virtue of* the existence of such an alternative possibility that Jones is morally responsible for what he does. I suggest that it is not enough for the flicker theorist to analyze the relevant range of cases in such a way as to identify an alternative possibility. Although this is surely a first step, it is not enough to establish the flicker of freedom view, because what needs to be shown is that these alternative possibilities *play a certain role* in the appropriate understanding of the cases. That is, it needs to be shown that these alternative possibilities *ground* our attributions of moral responsibility. And this is what I find puzzling and implausible.[20]

And Pereboom also endorses a similar requirement. He endorses Fischer's general point about the need for alternatives to be 'robust', and elaborates by proposing that any significant principle adverting to alternate possibilities should specify a necessary condition for moral responsibility which:

… plays a significant role in explaining why an agent is morally responsible. For if an agent is to be blameworthy for an action, it seems crucial that she could have done something to avoid being blameworthy – that she could have done something to get herself off the hook. If she is to be praiseworthy for an action, it seems important that she could have done something less admirable.[21]

It is evident, I think, that the kind of alternate possibilities I have insisted must always be present in any case of action will not meet the robustness criterion, as thus formulated by Fischer and Pereboom. That I might have entered the burning house through the window instead of the door is neither here nor there as far as my moral responsibility is concerned; that an addict might have injected heroin into his right arm rather than his left is certainly not the sort of thing that could render him morally blameworthy for what he does. The position offered might look as though it merely highlights alternate possibilities which are bewilderingly beside the point. I must now show, therefore, why this is not the case.

Essentially, the basis of my response to the worry about robustness is that the alternate possibilities I have highlighted are beside the point only in the context of the tradition which supposes that alternate possibilities matter to moral responsibility because they matter to the *fairness* of praising and blaming. In such a context, it would no doubt be right to demand of the defender of (PAP) that any alternate possibility she highlights is one which relates to praise and blame in the way Fischer and Pereboom suggest – e.g., if it is, for instance, a blameworthy action that is in question, that the alternate possibility show how the agent could have done something to 'get herself off the hook', to quote Pereboom. But on my view, alternate possibilities do not matter to moral responsibility in this way. They matter to moral responsibility not directly, but rather indirectly – they matter to moral responsibility because, and only because, they matter to agency. Determinism is inconsistent with moral responsibility not because it makes blame and praise unfair – but because it is inconsistent with the very existence of agents. And if there cannot be agents, there cannot be anything that has moral responsibility – for there cannot be anything – any entity – which controls and directs what occurs. There cannot be anything that anything is up to. In this context, there is no need to justify the relevance of an alternate possibility by showing directly how its existence relates to the blameworthiness or praiseworthiness of the imagined agent. One need

[20] Fischer (1994:140).
[21] Pereboom (2001: 1).

show only that it is plausible that unless *some* such alternate possibilities are presupposed we lose our grip entirely on the idea that we have an agent acting in the first place.

It is not true, of course, that we are morally responsible only for things which actions. We can be responsible also for omissions, for facts, for our beliefs and desires, etc. But it seems plausible that we can be responsible for these sorts of things only because it is presupposed that there are things we could have *done* to ensure that things had turned out differently. We can be responsible for an omission only if we could have acted; we can be responsible for a fact only if there is something we could have done (at least at *some* stage) to alter it; we can be responsible for a false belief only if we could (e.g) have investigated further or more carefully. If there could be no actions under determinism, it seems that there could be no moral responsibility either.

Once alternate possibilities are regarded as important for this reason, there is no need any longer for them to meet the condition imposed by Fischer and Pereboom on their relevance. For their relevance is not supposed any longer to be based on considerations having to do with fairness. It is supposed to be based on considerations to do with what it takes for an agent to exist and for an action to have occurred. What occurs in a case in which a Frankfurtian intervener intervenes is simply not an action on the part of the original agent – and so it remains true that there is an alternate possibility condition which she is able to meet – she was able not to have *acted* in the way that she did.

(v) *Leeway Incompatibilism and Source Incompatibilism*

The widespread perception that Frankfurt-style examples reveal that determinism and moral responsibility can co-exist, has generated a number of attempts to develop forms of incompatibilism rather different from the traditional sort – in the hope of capturing what it is that seems threatening about determinism in a different way. Some, for example, have distinguished between what they call 'leeway incompatibilism' and 'source (or 'causal history') incompatibilism'. The leeway incompatibilist, as characterised by Pereboom, holds that an action is free in the sense required for moral responsibility only if the agent could have done other than she actually did. The source, or causal history incompatibilist, on the other hand, holds that an action is free in the sense required for moral responsibility only if it is not produced by a deterministic process that traces back to causal factors beyond the agent's control.[22] Frankfurt-style cases, it is thought by many, may indeed show that leeway incompatibilism is incorrect. But they do not show that *source* incompatibilism is incorrect – since Frankfurt-style agents are typically such that their actions *are* produced by deterministic processes that trace back to causal factors beyond the agent's control.

For the Agency Incompatibilist, though, this way of looking at the matter separates two conditions which ought to be connected to one another. According to the Agency Incompatibilist, the leeway condition and the source condition are related. Agents are indeterministic initiators of chains of events and thus constitute the *sources* of such chains (in at least one of the senses relevant for moral responsibility) when and only when they *act* – i.e. exercise a power that is essentially *two-way* – and hence allows for leeway. That the power of agency has to be two-way (thus allowing for leeway) is connected tightly to the source condition – for it is the fact that the power is two-way which makes it the case that it is truly the *agent* (and not merely a set of events and states occurring or obtaining inside her) that brings about the bodily movement in the causing of which her action consists. That each actual movement or change brought about is something the agent needn't have brought about

[22] See Pereboom (2001: 2-3).

in quite the way or in quite the place, or at the very time she did in fact bring it about is crucial to the idea that the agent is indeed the *source* of what happens – for it is what ensures that deterministic chains cannot be traced back beyond her action to any conditions which preceded and necessitated it.

There are of course conceptions of sourcehood which are much richer and more demanding that any that is involved in the mere concept of an agent. Robert Kane, for example, speaks of wishing to retrieve a traditional sense for the term 'free will', in which it designates 'the power of agents to be the ultimate creators (or originators) and sustainers of their own ends or purposes'.[23] Galen Strawson clearly also has powers similar to those mentioned by Kane in mind when he speaks of our wish to be 'truly self-determined' where one can be truly self-determined 'only if one has somehow or other *determined how one is in such a way that one is truly responsible for how one is*'.[24] And Thomas Pink, like Kane, again stresses the centrality of the will, conceiving of it as a capacity for decision-making which is informed by practical reason, a capacity which he explicitly denies any non-human animal could possess.[25] Pereboom's version of hard incompatibilism and Smilansky's suggestion that moral responsibility must be an illusion both take such a conception of the type of freedom on which true moral responsibility depends for granted in arguing that since no such freedom does (or even could) exist, moral responsibility cannot truly exist either.[26] On the question whether we must have influence-busting freedoms of *this* sort in order to be morally responsible for anything, I remain agnostic at present; though I think my views are currently actually closer to compatibilist ways of thinking, than they are to incompatibilist ideas. It is incontrovertible that all of us get our principles and values from somewhere and that in many respects we are creatures of our upbringings and the various other types of social and cultural conditioning to which we are subject throughout our lives. Any conception of the conditions of moral responsibility which would require us somehow to formulate principles and to inculcate habits in ourselves in ways which float utterly free of any such societal underpinnings would evidently demand the impossible. But my inclination at that point is to say not that no one really bears any moral responsibility for anything, but rather to look for a conception of moral responsibility which might genuinely be reasonably be thought applicable to such beings as we evidently are. I think it unlikely that we would be forced thereby to retreat all the way to a position which allowed only for punishment for deterrence and prevention – which could find no place for a sensible conception of desert at all. And the important point on which I would continue to insist is that even if we cannot be the source of our actions in quite the way that, for example, Strawson supposes might be required for real, full-blooded moral responsibility, it does not follow that we cannot be their source in the weaker sense demanded by the Agency Incompatibilist - nor that our being such a source is not a much more evidently necessary condition of moral responsibility than its stronger cousin.

I hope I have done enough now to characterise Agency Incompatibilism, to give some sense of what might be the motivation for it, and to compare and contrast it with some of the other positions which are current in the literature. In the final third of this paper, I want to turn to attempt to meet some of the most obvious objections that are likely to be raised to the view I have described.

---

[23] Kane (1996: 4).
[24] Strawson (1986: 26).
[25] Pink (2004, Ch.2).
[26] See Pereboom (2001) and Smilansky (2000).

(vi) *Begging the Question*

In my work, I am sometimes accused of having 'begged the question' against the compatibilist in adopting the account of the nature of action that I do. But my reply to this is very short: it is that it is not clear at all why it is I who have begged the question in refusing the Causal Theory of Action rather than the compatibilist who has begged the question in favour of his own view, by adopting it. Moreover, it is my view, and not his, I believe, that has empirical support on its side, if the question is: what are actions according to our folk understanding? How do we conceive of them? The work to which I have already adverted in empirical psychology already provides much evidence, I believe, that the concept of an agent is the deep and complex concept I have suggested it is. It is, as I have already claimed, the causal theorist who is in danger of misrepresenting what folk psychology takes agency to involve.

Historically speaking, it seems to me fairly clear that the conception of action as involving events produced by prior constellations of further events and standing conditions owes its dominance of the modern era largely to the seventeenth century scientific revolution and the thinking about mechanism, the relation of whole to part, the dependence of macroscopic phenomena on the microscopic, etc., which that revolution brought in its train. Prior to that revolution, philosophers had mostly utilised an ontological scheme for thinking about action and activity that owed its main lineaments to Aristotle – a scheme in which the notion of a substance with distinctive powers of various sorts which were exercised in action played a central role, and in which the concept of an event can hardly be detected at all. These two contrasting ontological schemes are, I think, in genuine competition where animal agency is concerned, because the agent-based approach will deny that the exercises of two-way powers which it calls 'actions' can be safely conceived of properly in the terms permitted by the Causal Theory. And perhaps because the science that was triggered by that scientific revolution has been so successful by comparison with its Aristotelian predecessor, there is some reason to suppose that the ontological scheme that emerged from the seventeenth century revolution is preferable. But it should also be said that science has moved on a great deal since the seventeenth century, perhaps especially in the field of biology which is the most relevant of the traditional sciences for the understanding of animal agency. It may be that some of the scientific developments which have taken place since the seventeenth century – in for example, complexity theory, dynamic systems theory, etc., might provide a basis on which more Aristotelian-style thinking about agency might be successfully re-invigorated.[27] In the light of such developments, then, I think it would be rash to presume that the Causal Theory of Action has any rights to be regarded as the default option.

I do not think, then, that there is much plausibility in the charge that any questions have been begged by the Agency Incompatibilist. Her view of action is simply a genuine competitor to the Causal Theory and neither position can be charged with begging the question against the other. We simply have to assess each theory on its merits and decide which makes for the better approach to the phenomena in which we are interested.

(vii) *Determinism and Physics*

If I had to name what I regarded as the chief obstacle to a fair hearing for Agency Incompatibilism, it would be the following line of thought. The question whether determinism or indeterminism is true is an empirical question which has not yet been

[27] This has been suggested by e.g. Juarrero (1999) and Murphy and Brown (2007).

properly settled – and which, moreover, it is the job not of philosophers, but rather of physicists to decide. Determinism, that is, might yet turn out to be true – for all that we are frequently told that it is an increasingly unpopular position amongst those who are thought expert enough to be allowed a vote on the matter. But in that case, the claim that agency is inconsistent with determinism is bound to be hostage to fortune in a way that strikes most people as completely implausible. Surely that there are agents and that they are sometimes to be found performing actions is one of the things we *know*. But the Agency Incompatibilist appears unable to accept this claim. If, as she supposes, determinism is inconsistent with agency, and if determinism might yet turn out to be true, it seems to follow that agency might yet turn out to be an illusion. And that seems completely absurd. Perhaps we could just about accept that if determinism turned out to be true, we would then have to accept that there were in fact no *free* agents; perhaps we would even have to accept the disquieting idea that moral responsibility of any full-blooded sort would have to be relinquished.[28] But we surely cannot accept that there would not be any agents or any actions under such conditions.

At this point, though, it is useful to distinguish between the following two claims:

P1: The question whether determinism is true is a question which can only be answered by physics.

P2: The question whether determinism is true is a question which may (one day) be settled by physics.

P1 conceives of the question whether determinism is true as *essentially* a physicists' question – it is physicists who are said to have the authority to tell us whether or not determinism is true – *and no one else does*. P2, on the other hand, merely concedes that it is possible that physicists will one day show that the universe is entirely deterministic. I accept P2 – so much allowance has to be made for the possibility of future surprising and extraordinary discoveries of which we can currently have no conception. But I do not accept P1 – and it is P1, as I shall shortly argue, that would be required to show that Agency Incompatibilism was an untenable position.

Why do I deny P1? In order to understand why, it will help to return to the definition of determinism with which we began. Recall that determinism is the thesis that "for any given time, a complete statement of the (temporally genuine or non-relational) facts about that time, together with a complete statement of the laws of nature, entails every truth as to what happens after that time". But that means that determinism is an entirely *general* claim, not merely about the physical facts, but about *all* the 'temporally genuine' facts, whether these can be captured by means of the concepts of physics or not. Those facts, on the face of it, include, for example, the biological, psychological, sociological and economic facts – about which no physicist can be presumed to have any particular expertise – as well as a large number of simple contingencies which do not fall within the purview of any particular discipline, such as that my binoculars are broken; there are no hedgehogs on Mars; David Cameron is the current Prime Minister of the UK; and the bottle of sun-lotion on my desk is almost empty. If physicists are to have the right to tell us whether all these facts, which have, on the face of it, nothing very much at all to do with the sorts of things that physicists study, are entailed by prior facts and laws, then that will require an argument.

Now, of course many are happy to embrace the idea that there is bottom-up determination of the whole of reality by the physical, which would imply that if the physical

[28] Though Fischer has suggested that perhaps this, too, is a bridge too far: "Our fundamental nature as free, morally responsible agents should not depend on whether the pertinent regularities identified by the physicists have associated with them (objective) probabilities of 100% (causal determinism) or, say, 98% (causal indeterminism)" (2006: 5).

realm operated deterministically, the whole of reality would have to operate deterministically too. But whether or not that doctrine is *true* is not, at the moment, the point at issue. The point at issue is whether the claim that there is such bottom-up determination is *itself* a doctrine of physics. And I claim that it is not. It is quite plainly a controversial doctrine of *metaphysics* – a doctrine which may be disputed, therefore, using the tools at the disposal not only of physicists but also of metaphysicians. P1 embodies the metaphysical doctrine which Nancy Cartwright has called *fundamentalism* – a doctrine which espouses a vision of the universe as entirely dependent for its progression from one state to the next on the laws and principles of the single science we call physics. In some moods, I confess, such fundamentalism can seem natural. But it is in fact an extraordinarily strong thesis. It seems to imply, for instance, that our everyday supposition that the world sometimes develops over time in ways which are deeply affected by such things as stock market crashes, revolutions, speeches, technological developments, by matters, in other words, that fall well outside the purview of anything any physicist ever studies – must be mistaken. Are we really supposed to believe that all these sorts of occurrences, where they have happened, are somehow no more than the epiphenomenal by-products of the initial conditions and physical laws? – that they have not been essentially affected by altogether higher-level forms of causal influence? That purely physical laws have *allowed for* the existence of such things as stock market crashes must be agreed. But that they have *dictated them* seems preposterous. And it is for this reason that I am inclined to deny P1.

Of course, the acceptance of the weaker P2 implies that it is epistemically possible that one day, a physicist *may* come along and reveal that determinism is true. And if that were to happen, and if the physics were incontrovertible, then I should at that point have to withdraw the claim that agency is essentially an indeterministic phenomenon, involving the exercise at the time of action of two-way powers. But the bare possibility of such an eventuality does not justify *now* the rejection of the thesis. What I claim is only that we know nothing at present which suggests that Agency Incompatibilism could not be true.

There will be those who doubt this. In the final section of the paper, I want to make some headway in addressing the worries of these doubters. What these doubters will believe is something like this: the account which the Agency Incompatibilist offers of the nature of agency involves *agents* and the mysterious related notion of *agent causation*. Even if we have certain worries about the Causal Theory of Action, they will think, it does not, at any rate, involve any commitment to such preposterous entities and relations as these. Can Agency Incompatibilism be absolved of the charge that it invokes a phenomenon that it is impossible to make unmysterious?

(viii) *The Alleged Impossibility of Agent Causation*

I have spoken of *agents* bringing about movements and changes in their own bodies. But what are these agents? Am I committed to spooky Cartesian souls with the power to make extraordinary interventions into the course of nature? The answer is that I am by no means so committed. My agents are *animals* – non-spooky individuals made of physical stuff of whose existence we should be in no doubt whatever. When I claim that agents may bring about movements and changes in their own bodies, what I mean is simply that *animals* may do this. All the agents I think we know of are animal agents, and I am sceptical about the coherence of any conception we might think we have of agents of another kind.

This may allay one kind of worry only to raise another, however. We are used to the thought that when a large entity (such as an animal) brings about movements and changes in

its parts, that is generally traceable to a causal influence that some of its *parts* had on other parts. We might, for example, speak loosely of a computer opening a program, and thereby bringing about changes in its innards, on its screen, and so on. But these changes, we think, are simply brought about by *other* changes in the computer's innards and parts – and those by further such changes, and so on. There is no sense in which *the computer itself* is needed in addition to explain what happens – events occurring within it and states of it and its parts seem to do all the explanatory work in a case like this. It is very natural, then, to think that causation by whole must always reduce to causation by parts – and that this must be as true of the animal kingdom as it is elsewhere.

The denial that this is so is crucial to Agency Incompatibilism. The Agency Incompatibilist must, I think, maintain that it is possible for a whole animal to bring about effects in its parts – effects on its own neurons, for instance, which in turn can bring about effects on muscles and thereby on limbs, digits, etc – in such a way that this influence of whole on part does not simply reduce to the influence of some or other *parts* on part. But can we make any sense of this idea? My answer is that we can make sense of it, but that in order to do so, hard work needs to be done to resist ways of thinking about reality which make it seem that such top-down influence must be impossible.

Kim summarises one powerful thought which appears to stand in the way of making sense of the idea of top-down or downward causation:

the difficulties [with downward causation] essentially boil down to the following single argument. If an emergent, M, emerges from basal conditions C, why can't C displace M as a cause of any putative effect of M? Why doesn't C do all the work in bringing about the putative effect of M and suffice as an explanation of why the effect occurred?[29]

To return a satisfactory answer to this question, I think we need to think about the concept of *coincidence*. In order for complex phenomena such as actions to occur, an enormous number of things have to occur *together*. If I am to type this paragraph correctly, for instance, the parts of my brain which make available a knowledge of the English language will have to be engaged. My motor system will need to arrange for each of eight fingers to be over the right keys at the right time. I will need to remain upright and balanced in my chair. I need to keep my eyes focused on the screen. I need to be thinking about what I want to say and those thoughts need to feed through in the right way to my motor system. And so on. And so we need an explanation not just of how each of the mechanisms subserving each separate ability is able to function, but also of how it is that all these phenomena are enabled to happen simultaneously and/or in the right order. It cannot be permitted, that is, to be a sheer coincidence that all the necessary activity occurs together – we need an explanation of how the requisite phenomena are *orchestrated*.

It is here, in the explanation of what would otherwise have to count as coincidence in nature that I think we will find in general that we need to raise our eyes from the 'basal conditions' mentioned by Kim. An explanation of basal conditions in terms of *prior* basal conditions is all very well – but it will never serve to answer the question how it is that all the basal phenomena involved in the production of complex effects have been enabled to occur *together* – since the co-occurrence of the prior basal conditions which are causally sufficient for the relevant complex effects may look *equally* coincidental looked at merely from the viewpoint provided by the lower level. The explanation of one giant coincidence in terms merely of another, prior one does not serve to resolve the coincidence – it does not serve to show us why it is not a giant cosmic accident that the relevant states and events have all

[29] Kim (2000: 318)

managed to occur at the same time. But this is one of the things that we need to have causally explained. And this is not merely a point about what human beings require for the purposes of elucidation and illumination. It is a point about something that needs to be supplied in the causal metaphysics – the requisite phenomena need to be *made to happen* at the right time. It is, I think, in reflection on the sorts of resources we might need in order to help us understand how the co-occurrences essential to many sorts of complex phenomena are to be provided for that top-down causation may find a home.

In *Metaphysics* VI 3, Aristotle considers the case of a man who dies at the hands of ruffians because he goes out for a drink to the well at the wrong moment, thus arriving at the well at the same time as the ruffians.[30] Perhaps we could give a causally sufficient condition for the man's being at the well in basic physical terms, and perhaps we could give a similar sufficient condition for the ruffians' presence at that time also. But for all that, it may still be a coincidence that they arrived at the well together, an unfortunate piece of happenstance. What this example shows is that there may be a question left to answer even when all the causally sufficient conditions are in – a question which may or may not have an answer. In Aristotle's case, as he specifies it, the question has no answer – there is no *further* explanation to be had of why the man and the ruffians arrived at the well together (although there *could* have been such an explanation – it could, for example, have been the case that an accomplice of the ruffians had suggested to the man that he go out to the well at a pre-arranged time). But in the case of the complex phenomena which must occur together or in the right order if actions are to happen, we cannot tolerate the thought that there is no such explanation. It cannot merely be a *coincidence*, for example, that the complex synchronous arrangements needed to sustain purposive activity in a given animal have managed to occur together. There must be a causal explanation of what has permitted that to happen. And that explanation, it is plausible to think, must, for a set of mechanisms and processes at any given level, involve appeal to a mechanism or system at some higher level in a hierarchy – a mechanism or system whose job it is to ensure that the sub-systems which are parts of it operate in harmony one with another. This is, in general, the schematic answer to Kim's question. C doesn't suffice (entirely) as an explanation of why M occurred, if C is itself a complex conjunction of conditions whose co-occurrence (or occurrence in a given order) itself needs explaining. And no *prior* set of conditions C-1 will suffice either as an explanation of such a complex conjunction which answers all the causal questions we need answered. For any such prior C-1 which is complex enough to explain a complex conjunction will *itself* likely be a similar complex conjunction. And we will never have an explanation of why any of these complex conjunctions fails to be a giant coincidence unless we look to higher level agencies and systems whose function it is to supervise the lower-level ones in order to ensure and arrange for the wanted simultaneities.

Is the supervenience of higher-level phenomena on lower-level ones challenged by the suggestion that we might need sometimes need to look to higher-level phenomena to explain causally the co-occurrence of effects at the lower level? I think the answer to this is 'no' – it could remain true, for all I have said, for instance, that any two systems identical in all their physical properties must share all their higher-level properties too – and that any change in a higher-level property must always be accompanied by some change in a lower-level one. But this does not imply that lower-level change is always what *causally explains* change at the higher level. For all supervenience dictates, it could, in some cases, be the other way about. It helps, I think, to understand how this might be possible to reflect on the fact that supervenience is a thesis which generally relates to one another properties or states of affairs or facts that obtain at the same instant in time. In order to think of the relationship

[30] Aristotle (1984), 1027b1-16.

between higher level conditions and the lower level conditions which the supervenience thesis alleges are always constitutively sufficient for them, therefore, we are often thinking of the world in a kind of instantaneous, freeze-frame, snapshot view. And there is something deeply misleading about this snapshot view, I think. It makes it seem as though an entity which exists only for an instant could do causal duty for one which is a persisting object. At the higher level, for instance, we might speak of 'the property of believing that *p* at *t*' supervening on a momentary neural arrangement which also exists 'at *t*'. But we have to remember that an instantaneous molecular arrangement does not in fact, of course, guarantee, by itself, anyone's believing anything. Someone's believing something is a *persisting* state – and so its existence requires that the molecular arrangement which exists at *t* either maintain itself, or manages to be succeeded in time by a whole series of appropriate further molecular arrangements. And the causation of the right sort of succession of arrangements may be something that can only be understood by abstracting away from individual neurons and their interactions. It might be for instance, that the set of neurons essential to the maintenance of a given belief changes over time so that there is no particular continuity to discern provided we remain focused at the neural level. There may be a systemic imperative to retain the belief which is totally invisible from the neural level – and so which operates as an effective and genuinely top-down constraint on the sequence of 'snapshot' states of the brain in terms of which the formulation of the supervenience thesis encourages us to think.

My general suggestion, then (though evidently it really needs more elaboration than I am able to provide here), is that there is reason to think that we can make sense of the idea of top-down causation – and therefore reason to think that we can make sense also of the related idea that there may be top-down effects operating in hierarchically organised entities such as animals. These top-down influences may be such as to justify the thought that entities to be found higher in the hierarchy might dominate in various ways entities to be found lower down – e.g. systems dominate organs, organs dominate tissues, tissues dominate cells, etc. And an animal is no more than the top-level system in the animalian hierarchy – the entity which dominates all sub-animalian systems and ensures their integration and orchestration in the services of the animal's overall ends and purposes by means of the crucial, two-way power of agency. By this means, I believe, agency can be found a naturalistically respectable place in the inventory of powers we need to recognise in nature. It is, of course, a hugely important and in many ways distinctive power – but it can be fitted into a picture according to which it emerges, just as the powers proper to organs, tissues, cells, etc emerge, alongside the evolution of the hierarchically organised systems which constitute biological life forms. Agent causationism is, I think, often judged untenable because what is envisaged is a link between agent and event that occurs at the very beginning of some causal chain that results in a bodily movement – an input that is prior to some series of purely event-causal neurological connections. This is indeed untenable. But it is not the only option. We can understand agent causation not as *prior* causation of a chain of merely physical occurrences by some mysterious agent-causal impetus, but as *top down* control of lower-level occurrences by the top-level system in a biological hierarchy – the animal itself.

(ix) *Conclusion*

I believe Agency Incompatibilism to be a coherent and naturalistically respectable version of libertarianism which avoids many of the main problems both of its event-causal rivals and the standard agent-causationist alternatives. It offers the promise of being able to vindicate the libertarian intuition that a world in which things merely follow on from other things entirely deterministically cannot support moral responsibility, while being able to accommodate Frankfurt-style examples, animal agency and many of the best of the compatibilist's

intuitions (e.g. the conviction that the notion of an agent untouched by causal influence is not the notion we want in order to characterise the way in which the agent must be the source of her actions). There remains, no doubt, much work to be done in defending and elaborating the position which is in its relative infancy compared with most of its rivals. But I think it has great promise – and it will be enormously interesting over the next few years to attempt to develop in full detail the metaphysical framework of top down causation, two-way power and process ontology that I believe is required fully to support it.

References

Alvarez, Maria (2009), 'Actions, Thought Experiments and the 'Principle of Alternate Possibilities'', *Australasian Journal of Philosophy* 87: 61-82.

Aristotle (1984), *Metaphysics* in J. Barnes (ed.) *The Complete Works of Aristotle* Vol. 2 (Princeton: Princeton University Press).

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

Carey, S (1985), *Conceptual Change in Childhood* (Cambridge, MA: MIT Press).

Davidson, D. (1973). 'Freedom to act'. In: Honderich (ed.) Essays on Freedom of Action. London: Routledge and Kegan Paul, 137–56. Repr. in Davidson (1980): 63–81.

Fischer, John Martin (1994), *The Metaphysics of Free Will* (Oxford: Blackwell).

Fischer, John Martin (2006), *My Way* (Oxford: Oxford University Press).

Fischer, John Martin and Ravizza, Mark (1998), *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press).

Frankfurt, Harry (1969), 'Alternate Possibilities and Moral Responsibility', *Journal of Philosophy*, 89; repr.in Frankfurt (1988), 1-10.

Frankfurt, Harry (1982), 'The Importance of What we Care About', *Synthese* 53: 257-72, repr. in Frankfurt (1988): 80-94.

Frankfurt, Harry (1988), *The Importance of What We Care About: Philosophical Essays* (Cambridge and New York: Cambridge University Press).

Gelman, R., (1990), 'First Principles Organize Attention to and Learning About Relevant Data: Number and the Animate-Inanimate Distinction as Examples', *Cognitive Science* 14: 79-106.

Gelman, R., Durgin, F. and Kaufman, L. (1995), 'Distinguishing between animates and inanimates: not by motion alone', in Sperber, Premack and Premack (eds.) (1995): 151-84.

Gergely, G. and Csibra, G. (2003), 'Teleological reasoning in infancy: the naïve theory of rational action', *Trends in Cognitive Sciences*, 7: 287-92.

Haji, Ishtiyaque (1998), *Moral Appraisability* (Oxford: Oxford University Press).

Hirschfeld, L.A. and Gelman S.A (eds.) (1994). *Mapping the Mind* (Cambridge: Cambridge University Press).

Juarrero, A. (1999), *Dynamics in Action: Intentional Behavior as a Complex System* (Cambridge: MIT Press).

Kane, Robert (1996), *The Significance of Free Will* (New York: Oxford University Press).

Nonreductive Physicalism (Berlin and New York: Walter de Gruyter): 119-38.

Kim, Jaegwon (2000), 'Making Sense of Downward Causation', in Peter Bøgh Andersen, Claus Emmeche, Niels Ole Finneman and Peder Voetmann Christiansen (eds.), *Downward Causation* (Aarhus University Press: Aarhus).

Leslie, Alan M. (1994), 'ToMM, ToBY and Agency: Core architecture and domain specificity', in Hirschfeld and Gelman (eds.) (1994), 119-48.

Murphy, Nancy, and Brown, Warren S. (2007), *Did My Neurons Make Me Do It?* (Oxford: Oxford University Press).

Nelkin, Dana (2008), 'Responsibility and Reason: Defending an Asymmetrical View', *Pacific Philosophical Quarterly* 89: 417-515.

Nichols, Shaun (2004), 'The Folk Psychology of Free Will: Fits and Starts', *Mind and Language* 19: 473-502.

O'Shaughnessy, Brian (1980), *The Will*, 2 vols. (Cambridge: Cambridge University Press).
Peacocke, Christopher (1979), *Holistic Explanation: Action, Space, Interpretation* (Oxford: Oxford University Press).
Pereboom, Derk (2001), *Living Without Free Will* (Cambridge: Cambridge University Press).
Pink, Thomas (2004), *Free Will: A Very Short Introduction* (Oxford: Oxford University Press).
Premack, David (1990), 'The infant's theory of self-propelled objects', *Cognition* 36: 1-16.
Smilansky, Saul (2000), *Free Will and Illusion* (Oxford: Oxford University Press).
Spelke, Elizabeth S., Phillips, Ann and Woodward, Amanda L., (1995), 'Infants' knowledge of object motion and human action', in Sperber, Premack and Premack (eds.) (1995): 44-78.
Sperber, D., Premack, D. and Premack A.J. (eds.) (1995). *Causal Cognition: A Multidisciplinary Debate* (Oxford: Oxford University Press).
Steward, Helen (1997), *The Ontology of Mind* (Oxford: Oxford University Press).
Steward, Helen (2009a), Fairness, Agency and the Flicker of Freedom', *Nous* 43: 64-93.
Steward, Helen (2009b), 'Sub-Intentional Actions and the Over-Mentalization of Agency' in Constantine Sandis (ed.) *New Essays on the Explanation of Action* (New York: Palgrave Macmillan).
Steward, Helen (2012), *A Metaphysics for Freedom* (Oxford: Oxford University Press).
Strawson, Galen (1986), *Freedom and Belief*, (Oxford: Oxford University Press).
Van Inwagen, Peter (1983), *An Essay on Free Will* (Oxford, Oxford University Press).
Wolf, Susan (1990), *Freedom within Reason* (Oxford: Oxford University Press).