there is some defect in one or both of our two arguments. Either there is something wrong with our argument for the conclusion that metaphysical freedom is incompatible with determinism or there is something wrong with our argument for the conclusion that metaphysical freedom is incompatible with *in*determinism – or there is something wrong with both arguments. But which argument is wrong, and why? (Or are they both wrong?) I do not know. I think no one knows. That is why my title is, "The *Mystery* of Metaphysical Freedom." I believe I know, as surely as I know anything, that at least one of the two arguments contains a mistake. And yet, having thought very hard about the two arguments for almost thirty years, I confess myself unable to identify even a possible candidate for such a mistake. My *opinion* is that the first argument (the argument for the incompatibility of freedom and determinism) is essentially sound, and that there is, therefore, something wrong with the second argument (the argument for the incompatibility of freedom and indeterminism). But if you ask me *what* it is, I have to say that I am, as current American slang has it, absolutely clueless. Indeed the problem seems to me to be so evidently impossible of solution that I find very attractive a suggestion that has been made by Noam Chomsky (and which was developed by Colin McGinn in his recent book *The Problems of Philosophy*) that there is something about our biology, something about the ways of thinking that are "hardwired" into our brains, that renders it impossible for us human beings to dispel the mystery of metaphysical freedom. However this may be, I am certain that I cannot dispel the mystery, and I am certain that no one else has in fact done so.

42 The Agent as Cause

Timothy O'Connor

In the previous essay, Peter van Inwagen argues that "metaphysical freedom" is incompatible with a certain abstract picture of the world (commonly dubbed "determinism"), on which it evolves in strict accordance with physical laws, laws such that the state of the world at any given time ensures a unique outcome at any subsequent moment. I agree that the two are incompatible. But what, in positive terms, does the ordinary understanding of ourselves as intelligent beings who "freely" decide how we shall act require? Where do the "springs of action" lie for beings that truly enjoy "free will"? This is surprisingly difficult to answer with any confidence. A useful way of approaching this question is to consider the various ways we might modify determinism in order to accommodate free will.

The most economical change in the determinist's basic picture is to introduce a causal "loose fit" between those factors influencing my choice (such as

my beliefs and desires) and the choice itself. We might suppose, that is, that such factors *cause* my choice in an *in*deterministic way. To say that the causation involved is "indeterministic" is perhaps to say that the laws governing the evolution of the world through time (including that bit of the world which is me) are fundamentally statistical: they allow that (at least at various junctures) a range of alternatives are possible, though they will specify that certain of them are far more likely than others, in accordance with some measure of probability. Applying this general idea to the case of human choices, one might suppose that a free choice requires the following features: I have reasons to act in accordance with each of a range of options. In each case, my having those reasons gives me an objective (probabilistic) tendency to act accordingly. But whatever the relative probabilities of the alternatives, each of them is possible. And whichever of them occurs, the agent's having had a specific reason so to act will have been among the factors that caused it. Let us call this modification of the deterministic picture "causal indeterminism."

Would this be freedom? In my judgment, it would not. It is not enough that any of a range of possible actions are *open* to me to perform. I must have the right sort of *control* over the way the decision goes in a given case. And we may ask of the causal indeterminist, how is it up to me that, on this occasion, this one among two or more causally possible choices was made? I find myself with competing motivations – in my present case, a desire to watch a basketball game, a desire to play a game with my children, and a desire to finish this article – each of a particular "strength." On this occasion, we may suppose, the least probable outcome occurs. On other occasions, more probable outcomes occur. If I am truly acting freely, then presumably I in some way directly control or determine which outcome occurs on a given occasion. But in what does that control consist? The causal indeterminist does not have resources, it seems, to satisfactorily answer this question. Given a sufficiently large number of choices of a large number of people, the pattern of outcomes is likely to conform, more or less, to the statistical character of the underlying laws. There seems nothing more that one can say – in particular, nothing more one can say about the outcome of any particular choice. The indeterministic tendencies arising from my reasons confer a *kind* of control that is too "chancy" to ground significant responsibility. Indeed, it does not differ at all in *kind* from the control that would be had in a deterministic world; it merely introduces an element of "looseness" into its exercise. Given this added looseness, the future *is* open to alternative possibilities. But it remains unclear how I myself could be responsible (in part) for which of those alternatives is realized.

A dilemma is forming. Responsibility for our actions is inconsistent with the deterministic picture of the world. But it is also inconsistent with at least one straightforward kind of indeterministic picture, the kind that most directly carries into the sphere of human action the sort of indeterminism that many theorists believe operates at the level of fundamental physics. Indeed, a good many philosophers suppose that these two pictures (which we have labeled "causal determinism" and "causal indeterminism") exhaust the plausible alternatives. If all this is right, then the conclusion to be drawn is that free will is simply an

inconsistent notion. It's not that we just don't happen to have free will; rather, we don't have it because it simply can't be had.

One alternative to this unpalatable conclusion is that entertained by Peter van Inwagen, in his contribution to this volume. Perhaps, van Inwagen writes, "there is something about our biology, something about the ways of thinking that are 'hardwired' into our brains, that renders it impossible for us human beings to dispel the mystery of metaphysical freedom" (see p. 374). That is, though the notion of free will isn't truly inconsistent, its nature is "cognitively closed" to us. (After all, we have no reason to be confident that we are able, even "in principle," to grasp *every* difficult notion that, say, God grasps. And the history of philosophical reflection on the idea of freedom of will suggests that it has its subtleties.)

Well, there is certainly no *arguing* against this suggestion, absent the emergence of a stable consensus of opinion on the matter – rather unlikely at this stage of the game. But one may well distrust it on the general grounds that it counsels complacency. (And why stop at the notion of free will? Philosophers disagree over the correct understanding of most significant philosophical concepts.) Furthermore, once a philosopher takes this suggestion seriously, he may well be drawn into a deeper measure of skepticism about the notion of freedom of will than initially intended. Van Inwagen, for example, tells us that he is of the opinion that free will is *in*compatible with determinism. So he supposes that it must be *compatible* with *indeterminism*, even though he fails to see *which* sort of indeterminism will clearly do the trick. But if he and the rest of us are "hardwired" in some manner that precludes our coming to understand adequately the nature of free will, is it likely that we understand it sufficiently to know even *some* of its features? At any rate, the hypothesis ought to automatically undercut one's confidence in any highly *disputed* claims, such as van Inwagen's relative confidence in the thesis that free will is incompatible with determinism. (I note that Colin McGinn, whom van Inwagen cites in this connection, supposes that free will *can* be had under determinism, even though he "can't see how".)

Rather than embrace the despair and skepticism of the "cognitive closure" hypothesis, then, let us pick up the argument where we last left it, and see whether a "positive" solution to our problem is in the offing. I argued that, if my decisions to act are simply the indeterministic effects of my beliefs and desires, then they are not up to me. What more do we *want* to say about our decisions, that causal indeterminism leaves out?

Just this, it seems: that I myself freely and directly control the outcome, where "control" here (as everywhere) is evidently a *causal* notion. And the unsatisfactoriness of causal indeterminism suggests that we have to be rather literal about the referent of "I," in this context. If I do something freely, I cannot be thought of as simply an arena in which internal and external factors work together to bring about my action (whether or not these factors are thought to operate in a strictly deterministic fashion). Instead, we want to say with Roderick Chisholm that I am the "end-of-the-line" initiator of the resulting action. What we are after, that is, is a notion of a distinctively personal form of

causality (in the parlance of philosophers, "agent causation"), as against the broadly mechanistic form of causality ("event causation") that both the deterministic and causal indeterministic pictures represent as governing *all* forms of activity in nature without exception.

Many philosophers find this notion of "personal" or "agent" causation to be utterly mysterious, or downright incoherent. (Some of those philosophers will agree that it is natural to talk of "agent causation" when trying to articulate an understanding of free will, even though it is an incoherent idea. On their view, the term encapsulates the inconsistent strands in that notion.) Here is a simple reflection that fosters the sense of mystery. We often talk loosely of inanimate objects as causing certain things to happen. An example is the statement that Zimmerman's car knocked down the telephone pole. But it's clear that this does not perspicuously capture the metaphysics of the situation. It is instead simply shorthand for the assertion that the *movement* of Zimmerman's car (a car with a certain mass) caused the pole's falling down. It is, then, this *event* involving Zimmerman's car that brought about the effect, and not simply the car, *qua* enduring object. (No such effects emanate from his car when Zimmerman wisely decides to keep it parked in his garage.) The problem that many see with agent causation is that it rejects any expansion of "loose" talk of agents' causing things to happen into statements asserting that particular events *involving* those agents cause the effects in question. And that can seem mysterious: how can agents cause things to happen without its being true that they do so in virtue of certain features of themselves at the time? The agent is, after all, always an agent; yet he is not always causing some particular effect, such as deciding to complete an article on agent causation. Doesn't this force us to acknowledge that if the agent has decided to complete that article at one particular time, there must have been something *about him* at that time in virtue of which that effect was realized? And isn't that just to say that the *event* of the agent's having those distinguishing features, whatever they were, is what caused the decision?

This simple reflection is perhaps the deepest basis for philosophical suspicion about the notion of agent causation. However, I have come to suspect the suspicion and its various bases. In order to have a clear view of this matter, we need to reflect further on what is involved in our ordinary understanding of causation. Unfortunately, there is precious little agreement among philosophers about these matters. But the brief remarks I will make on this score at least have the advantage of representing a fairly commonsensical view of causation.

On the theory of causation I favor, objects are inherently active or dynamic. That is, they have causal capacities, and these are not "free-floating", but rather are linked to their intrinsic properties – those basic properties whose exact character it is the business of science to investigate.

In the more generally applicable case of *event* (or broadly mechanistic) causation, the *exercise* of such a capacity or tendency proceeds "as a matter of course": a thing's having, in the right circumstances, the capacity-grounding cluster of properties directly generates one of the effects within its range. (For indeterministic capacities, that effect will be but one of a range of *possible* ef-

fects; whereas in the deterministic case, there is only one possible outcome.)

The way that agent or personal causation differs from this mechanistic paradigm is in the way the relevant causal capacities are *exercised*. An agent's capacity to freely and directly control the outcome of his deliberation also requires underlying intrinsic properties which ground that capacity. (What sort of properties these might be is an interesting and in certain respects puzzling question, but it is at least partly empirical and not conceptual in nature. In any case, I shall not consider it here.) And no doubt the range of its operation is sharply circumscribed. For what is it, after all, that I directly act on, according to the agency theory? Myself – a complex system regulated by a host of stratified dynamic processes. I don't introduce events *ex nihilo*; (at best) I influence the direction of what is already going on within me. What is going on is a structured, dynamic situation open to some possibilities and not others. So the capacity is also circumscribed by physical and psychological factors at work within the agent while he deliberates. But (and here is the difference from the mechanistic paradigm) having the properties that subserve an *agent*-causal capacity does not suffice to bring about a particular effect (or even the occurrence of some effect or other within a range of possible effects); rather, it *enables* the agent to determine an effect (within the corresponding range). Whether, when, and how such a capacity will be exercised is freely determined by the agent.

That is the core metaphysical difference between the two causal paradigms. But we have yet to discuss how prior desires, intentions, and beliefs (more simply, "reasons") may explain such agent-causal activity. I suggest that we think of the agent's immediate effect as an action-triggering state of *intention* (which endures throughout the action and guides its completion). The content of that intention, in part, is that I act here and now in a particular sort of way. But another aspect of that intention, in my view, is that an action of a specific sort be performed *for certain reasons* the agent had at the time. (After a brief deliberation, I formed the intention to continue to type these words *in order to get the editors of this volume off my back*.) And the basis of the explanatory link lies precisely in this fact that the intention refers to the guiding reason. That is, the caused intention bears its explanation on its sleeve, so to speak. Had the agent generated a different intention, it would have been done (in most cases) for a different reason, to which reason the content of the intention itself would have referred. And if the agent had *several* reasons for performing a particular action, the reason(s) that *actually* moved the agent to act, again, would be reflected in the content of the intention. (None of this is to suggest that determining this content, in retrospect, is always easy. Clearly, I can be mistaken about my own reasons for acting.)

Some say that this account of the explanatory nature of reasons cannot be right: we can simply see that any undetermined instance of agent causation would be random, since by hypothesis nothing causes it. (Even some proponents of agent causation have been worried about this, and have been led to posit infinite hierarchies of agent-causings.) But it is hard to credit this objection. Consider what is being demanded. Agent causation is a form of direct control over one's behavior *par excellence*. But this is held to be insufficient.

What is needed, it is argued, is some mechanism by virtue of which the agent controls this controlling. Put thus (though understandably it is not generally put in this way), its absurdity is evident. We needn't control our exercises of control. (For if we did, then wouldn't we also need yet another exercise of control, and so on?) On any coherent conception of human action, there is going to be a *basic* form of activity on which rests all control over less immediate effects. On the agent-causal picture, this basic activity is that of an agent's directly generating an intention to act in accordance with certain reasons.

Others have argued that the suggested account of explanations of free actions by reasons cannot be right, since the reasons to which one points in a given case won't explain why the agent acted as he did *rather than* in one of the other ways that were open to him (alternatives that by hypothesis remained open up to the very moment of choice). But while the issues involved here are subtler, this objection also fails. The objection assumes that adequately explaining an occurrence *ipso facto* involves explaining why that event occurred rather than any imaginable alternative. And this seems too strong a requirement. At bottom, explaining an occurrence involves uncovering the causal factor that generated it. In deterministic cases, where only one outcome is possible, such an explanation will also show why that event occurred rather than any other. But this should not blind us to the fact that the two targets of explanation are distinct: the simple *occurrence* itself and the *contrastive fact* that the outcome occurred rather than any other alternative. We need this distinction not just to understand human free agency, but to understand any indeterministic causal activity, including the apparently indeterministic mechanisms described by physical science. Whether (and in what circumstances) there can *also* be contrastive explanations of such indeterministic outcomes is a difficult question. But whatever we say here, there is little to recommend the claim that an occurrence that has been caused, though not uniquely determined, by some factor is thereby wholly inexplicable.

More might be said about the "nature of reasons" explanation on the picture just sketched, but I want to turn instead to the complaint that we've swung too far in the direction of freedom. In place of the diminished, freedom-less conception of human action entailed by the deterministic picture, we've substituted a rather god-like one: the agent selects from among reasons that are merely passively present before the agent as he deliberates, reasons that do not *move* the agent to act. Though rather implausible on the face of it, such a consequence is embraced by some advocates of agent causation. Chisholm, for example, compares agent causation with divine action:

> If we are responsible, and if what I have been trying to say is true, then we have a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved. In doing what we do, we cause certain events to happen, and nothing – or no one – causes us to cause those events to happen.[1]

But perhaps this is unnecessarily heroic. Though defenders of agent causation have generally insisted on a sharp divide between it and mechanistic causa-

tion, we may be able to move tentatively toward greater integration of the two. The goal is not to *reduce* agent causation, in the end, to an all-encompassing mechanistic paradigm, but rather to see how event-causal factors such as the possession of reasons to act may *shape* the distinctively agent-causal capacity. Two things, in particular, seem needed here – if not for all conceivable agents (including God and angels), then at least for human beings as we know them. First, our account should capture the way reasons (in some sense) *move* us to act as we do – and not as external pressures, but as *our* reasons, as our own internal tendencies to act to satisfy certain desires or aims. Secondly, the account should acknowledge that those reasons typically do not have "equal weight," so to speak. It is a truism that, given the structure of my preferences, stable intentions, and so forth, and the situation with which I am faced, I am often far more likely to act in one way rather than in any other. But how might we account for this, if not in terms of a relative tendency, on the part of reasons, to *produce* our actions?

In my view, this is the biggest obstacle to a clear understanding of what free will requires. What we need is a way to modify the traditional notion of a distinctively personal kind of causal capacity and to see it, not as utterly unfettered, but as one that comes "structured", in the sense of having built-in propensities to act (though ones that shift over time in accordance with the agent's changing preferences). But we must do so in such a way that it remains up to me to act on these tendencies or not, so that what I do is not simply the consequence of the vagaries of "chance-like" indeterministic activity, as may be true of microphysical quantum phenomena.

So, the task of harmonizing free and responsible human agency with a world that is fundamentally mechanistic in character remains unfinished. But perhaps we've seen enough to dispel much of the air of profound mystery that some profess to find on considering the very idea of metaphysical freedom.

Note

1 "Human Freedom and the Self," p. 362, this volume.