Philosophical Perspectives, 14, Action and Freedom, 2000

FREE WILL REMAINS A MYSTERY

The Eighth *Philosophical Perspectives* Lecture

Peter van Inwagen
The University of Notre Dame

This paper has two parts. In the first part, I concede an error in an argument I have given for the incompatibility of free will and determinism. I go on to show how to modify my argument so as to avoid this error, and conclude that the thesis that free will and determinism are compatible continues to be—to say the least—implausible. But if free will is incompatible with determinism, we are faced with a mystery, for free will undeniably exists, and it also seems to be incompatible with *in*determinism. That is to say: we are faced with a mystery if free will *is* incompatible with indeterminism. Perhaps it is not. The arguments for the incompatibility of free will and indeterminism are plausible and suggestive, but not watertight. And many philosophers are convinced that the theory of "agent causation" (or some specific development of it) shows that acts that are undetermined by past states of affairs can be free acts. But the philosophical enemies of the idea of agent causation are numerous and articulate. Opposition to the idea of agent causation has been based on one or the other of two convictions: that the concept of agent causation is incoherent, or that the reality of agent causation would be inconsistent with "naturalism" or "a scientific world-view." In the second part of this paper, I will defend the conclusion that the concept of agent causation is of no use to the philosopher who wants to maintain that free will and indeterminism are compatible. But I will not try to show that the concept of agent causation is incoherent or that the real existence of agent causation should be rejected for scientific reasons. I will assume—for the sake of argument—that agent causation is possible, and that it in fact exists. I will, however, present an argument for the conclusion that free will and indeterminism are incompatible even if our acts or their causal antecedents are products of agent causation. I see no way to respond to this argument. I conclude that free will remains a mystery—that is, that free will

undeniably exists and that there is a strong and unanswered *prima facie* case for its impossibility.

I

I have offered the following argument for the incompatibility of free will and determinism.[1] Let us read 'N*p*' as '*p* and no one has or ever had any choice about whether *p*'. We employ the following two inference rules

α $\Box p \vdash \mathrm{N}p$
β $\mathrm{N}p, \mathrm{N}(p \supset q) \vdash \mathrm{N}q$.

(The box, of course, represents necessity or truth in all possible worlds.) Let 'L' represent the conjunction of the laws of nature into a single proposition. Let 'P_0' represent the proposition that describes the state of the world at some time in the remote past. Let 'P' represent any true proposition. The following statement, proposition (1), is a consequence of determinism:

(1) $\Box((\mathrm{P}_0 \,\&\, \mathrm{L}) \supset \mathrm{P})$.

We now argue,

(2)	$\Box((\mathrm{P}_0 \supset (\mathrm{L} \supset \mathrm{P}))$	1, standard logic
(3)	$\mathrm{N}((\mathrm{P}_0 \supset (\mathrm{L} \supset \mathrm{P}))$	2, α
(4)	NP_0	Premise
(5)	$\mathrm{N}(\mathrm{L} \supset \mathrm{P})$	3, 4, β
(6)	NL	Premise
(7)	NP	5, 6, β.

Since the two premises are obviously true—no one has any choice about the past; no one has any choice about the laws of nature—, (7) follows from (1) if the two rules of inference are valid.[2] And from this it follows that if determinism is true, no one has any choice about anything.

Are the two rules of inference valid? Rule α obviously is, whatever Descartes would have us believe about God. The question of the soundness of the argument comes down to the question whether β is valid. And, although β does not, perhaps, share the "luminous evidence" of α, it nevertheless seems pretty plausible. One way to appreciate its plausibility is to think in terms of regions of logical space. By logical space, I mean a space whose points are possible worlds. (Distances between points correspond to the "distances" that figure in a Lewis-Stalnaker semantics for counterfactual conditionals; areas or volumes represent probabilities[3].) Consider Figure 1.

Suppose Alice is inside *p* and has no choice about that; suppose she is also inside the region that corresponds to the material conditional whose antecedent

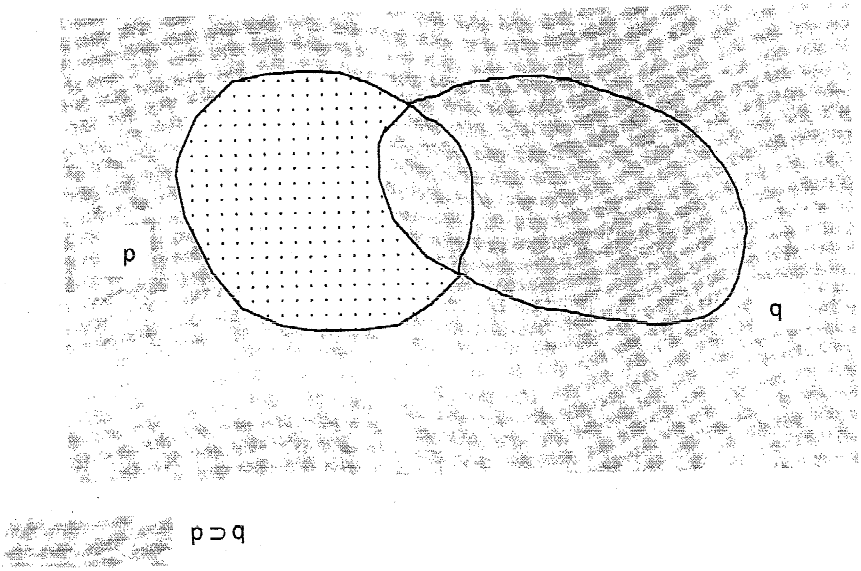

Figure 1.

is *p* and whose consequent is *q* (the heavily shaded region)—and has no choice about *that*. Alice will, of course, be inside the intersection of *p* and *q*, and hence inside *q*.[4] Has she any choice about that? It would seem not. As an aid to our intuitions, let us think of the regions displayed in the diagram as physical regions. Examination of the diagram shows that any way out of *q*—any escape route from *q*, so to speak—will either take Alice out of *p* or out of the shaded region. Therefore, *because* Alice has no way out of *p* and no way out of the shaded region ($p \supset q$), she has no way out of *q*. To be inside a region and to have no way out of it is to be inside that region and to have no choice about whether one is inside it. Rule β, therefore, would seem to be valid. This intuitive, diagrammatic argument is very plausible, and at one time I found it, or something very like it, cogent. Unfortunately, as any student of geometry knows, figures can be misleading, since a figure may have unintended special features that correspond to unwarranted assumptions. And this must be so in the present case, owing to the fact that McKay and Johnson have discovered what is undeniably a counterexample to β.[5]

McKay and Johnson begin by noting that α and β together imply the rule of inference that Michael Slote has called Agglomeration:

$$Np, Nq \vdash N(p \,\&\, q).$$

(To show this, assume N*p* and N*q*. The next line of the proof is '$\Box(p \supset (q \supset (p \,\&\, q)))$'. The proof proceeds by obvious applications of α and β.) Rule α is

obviously correct. To show β invalid, therefore, it suffices to produce a counterexample to Agglomeration. McKay and Johnson's counterexample to Agglomeration is as follows.

> Suppose I have a coin that was not tossed yesterday. Suppose, however, that I was able to toss it yesterday and that no one else was. Suppose that if I had tossed it, it might have landed "heads" and it might have landed "tails" and it would have landed in one way or the other (it's false that it might have landed on edge, it's false that a bird might have plucked it out of the air...), but I should have had no choice about which face it would have displayed. It seems that
>
> N The coin did not land "heads" yesterday
>
> N The coin did not land "tails" yesterday
>
> are both true—for if I had tossed the coin, I should have had no choice about whether the tossed coin satisfied the description 'did not land "heads"', and I should have had no choice about whether the tossed coin satisfied the description 'did not land "tails".' But
>
> N (The coin did not land "heads" yesterday & the coin did not land "tails" yesterday)
>
> is false—for I did have a choice about the truth value of the (in fact true) conjunctive proposition *The coin did not land "heads" yesterday and the coin did not land "tails" yesterday*, since I was able to toss the coin and, if I had exercised this ability, this conjunctive proposition would have been false.

The case imagined is, as I said, undeniably a counterexample to Agglomeration. Agglomeration is therefore invalid, and the invalidity of β follows from the invalidity of Agglomeration. Our diagrammatic argument for the validity of β therefore misled us. But what is wrong with it?

We may note that a similar intuitive, diagrammatic argument could have been adduced in support of Agglomeration. Imagine two intersecting regions, p and q. Their region of overlap is, of course, their conjunction. Suppose one is inside p and has no way out of p; and imagine that one is inside q and has no way out of q. One will then be inside p & q; but does it follow that one has no way out of p & q? Inspection of the simple diagram that represents this situation shows that any way out of p & q must either be a way out of p or a way out of q. What is wrong with *this* argument?

To answer this question, we must examine the concept of "having a way out of a region of logical space." Suppose we know what is meant by "having access to" a region of logical space. (A region of logical space corresponds to a proposition, or to a set containing a proposition and all and only those propositions necessarily equivalent to it. To have access to a region of logical space is to be able to ensure the truth of the proposition that corresponds to that region, or to be able to ensure that that region contains the actual world. If one is

inside a region one *ipso facto* has access to that region. If one has access to *p*, one *ipso facto* has access to the regions of which *p* is a subset—to the "superregions" of *p*.) To have a way out of a region *p* of logical space that one is inside is then defined as follows: to have access to some region that does not overlap *p*—or to have the ability to ensure that the proposition that corresponds to *p* is false. Now consider Figure 2.

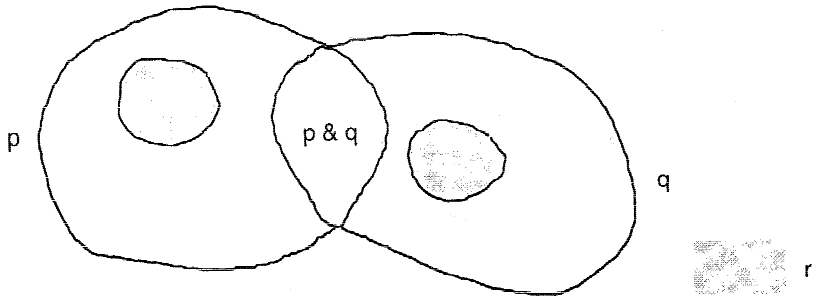

Figure 2.

Suppose I am "inside" the region *p* & *q*. Suppose I have access to and only to the following regions: (a) *p* & *q* and the other regions I am inside, and (b) *r* and the superregions of *r*. ("But what about the subregions of *r*?" From the fact that one has access to a certain region of logical space, it does not follow that one has access to any of its proper subregions. I may, for example, be able to ensure that the dart hit the board, but unable to ensure with respect to any proper part of the board that it hit that proper part.) It follows from these suppositions that I am inside *p* and have no way out of *p*—for every region to which I have access overlaps *p*. (And, of course, the same holds for *q*: every region to which I have access overlaps *q*.) But I do have a way out of *p* & *q*, for I have access to a region—*r*—that does not overlap *p* & *q*. (It is not essential to the example that *r* be a non-connected region. It might have been "horseshoe-shaped" or a "ring." What is essential is that *r* overlap *p* and overlap *q* and not overlap *p* & *q*.)

If one thinks about the issues raised by McKay and Johnson's counterexample in terms of diagrams of logical space, it is easy enough to construct a counterexample to β itself (at least in the sense in which Figure 2 represents a counterexample to Agglomeration).[6] Here is a simple counterexample to β. Consider three regions of logical space, related to one another as in Figure 3:

Suppose I am inside *p* and inside $p \supset q$. (Or, what is the same thing, suppose I am inside *p* & *q*.) Suppose I have access to and only to the following regions: (a) the regions I am inside, and (b) *r* and its superregions. Then I have no way out of *p* (every region to which I have access overlaps *p*) and no way out of $p \supset q$ (every region to which I have access overlaps $p \supset q$), but I have a way out of *q*, for I have access to a region—*r*—that does not overlap *q*.

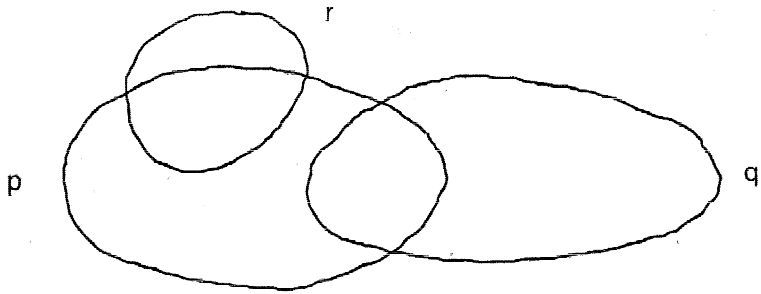

Figure 3.

How did Figure 1 and the intuitive argument based on it mislead us? The answer is simple. The informal argument invited us to think of "having a way out of a region" as something like having available a path or line leading from a particular point inside that region *to a particular point* outside that region. (Recall our use of the term "escape route.") That, after all, is what it is normally like to have a way out of a region of physical space, and our intuitive grasp of any sort of space is mainly by way of analogy with physical space. But if we exercise our imaginations, we can think of ways in which one might have an ability to change one's position in physical space that is entirely different from the ability to follow a path that leads to a given point. We might for example suppose that one can bring it about that one changes one's position in space without moving—by magic, perhaps—, and that when one changes one's position by this means, one might arrive at *any* of the points that make up some extended region.

Now consider once more Alice and Figure 1 (but add to Figure 1 a region *r* that is related to *p* and *q* just as *r* is related to *p* and *q* in Figure 3). Our intuitive argument for the conclusion that a way out of *q* must either be a way out of *p* or a way out of $p \supset q$ (the shaded region) was this:

> As an aid to our intuitions, let us think of the regions displayed in the diagram as physical regions. Any way out of *q*—any escape route from *q*, so to speak—will either take Alice out of *p* or out of the shaded region.

As long as Alice moves by following a continuous path through space (an "escape route"), this is correct: any continuous path that leaves both *p* and the shaded region must leave *q*. But suppose that although Alice has no way of crossing any of the boundaries shown in the diagram by following a continuous path through space, she has a single magical resource: a magical lamp such that if she rubs it, the Slave of the Lamp will instantaneously translate her to a randomly chosen point inside the region *r*. Has Alice a way out of *p*? Has she a way out of $p \supset q$? The answers to these questions, perhaps, depend on how

one defines 'a way out'. But if we define 'a way out' in a way parallel to our definition of 'a way out of a region of logical space', i.e.,

> If one is inside a region of space *r*, one has a way out of *r* just in the case that one is able to ensure that one is inside a region that does not overlap *r*,

the answer to both questions is No: she has no "way out" of either of these regions. But Alice *does* have a way out of *q*: rubbing the lamp constitutes a way out of *q*, for rubbing the lamp will ensure that she is not in *q*. Our intuitions about physical space therefore misled us. As the world is, the only way to leave a region of physical space is to follow a continuous path out of that region, and our intuitions reflect this fact. Our diagrams of logical space are, of course, drawn in physical space and the diagrams therefore invite us to think of one's having access to a region *r* of logical space (a false proposition, a region not containing the actual world) in terms of one's ability to move along a line drawn from the point in the diagram that represents the actual world to some point inside the section of the diagram that represents *r*. Our "diagrammatic" argument misled us into thinking that there could be no counterexample to β (or to Agglomeration) because nothing in the concept of "access to a region of logical space" corresponds to the "continuous path" requirement that the real world imposes on our intuitions about "access to a region of physical space." A continuous path through physical space terminates in a single point, not in an extended region. To "have access" to an extended region of physical space is therefore (normally) to have access to one or more of the points that make up that region. To have access to a region of logical space, however, is in no possible case to have access to a point in logical space (a single possible world): Since one's power to direct the course of events is limited, from the fact that one is able to ensure that *some* possible world in which, say, the coin is tossed is actual it does not follow that one is able to ensure with respect to any given world in which the coin is tossed that *that* world is actual. And of course, one never is able to ensure this; if one were one would not only be able to ensure that a tossed coin land on one particular face, but one would be able to determine the truth-value of every contingent proposition.

Our definition of 'N*p*' was '*p* and no one has or ever had any choice about whether *p*'. The definiens is equivalent to

> *p* and every region of logical space to which anyone has, or ever had, access overlaps *p*.

Why? Well, suppose that *p*, and that I did not do but was able to do X (and was able to do nothing else that was relevant to the truth-value of *p*), and that if I had done X, *p* might have been true and might have been false. It seems wrong in that case to say that I had a choice about the truth-value of *p*. If, for example, the coin was untossed and I was able to toss it, and if I had tossed it, it

might have fallen "heads" and might not have fallen "heads," it is wrong to say that I had a choice about the truth value of the (true) proposition that the coin did not land "heads." (This point is the essence of the McKay-Johnson counterexample to Agglomeration.) Now if it were important that the coin have landed "heads" (if someone's life depended on its landing "heads," say), there would be something wrong with my defending my failure to toss the coin by saying, "Look, the coin *didn't* land heads, and I didn't have any choice about that." And it is perhaps intuitively plausible to suppose that if *p* and if I had no choice about whether *p*, then I cannot properly be held morally responsible for *p*. But I don't think that this consideration has any tendency to show that I had a choice about how the coin fell. If I did offer the imagined lame excuse, the proper response would not be, "You did too have a choice about whether the coin landed 'heads"'; it would rather be, "You had a choice about whether the coin was tossed, and if you had tossed it, it might have landed 'heads.' What you are to blame for is not doing your best to bring it about that the coin landed 'heads."' In sum, if *p* is true proposition, having a choice about the truth-value of *p* implies being able to *ensure* that *p* is false.[7] And, as we have seen, the following is possible: *p* is true and no one is able to ensure that *p* is false; the conditional whose antecedent is *p* and whose consequent is *q* is true, and no one is able to ensure that that conditional is false; someone is able to ensure that *q* is false.

McKay and Johnson are therefore right. Rule β is invalid, and my argument for the incompatibility of free will and determinism is invalid.

This, of course, does not imply that free will and determinism are compatible, or that there is no plausible argument for the incompatibility of free will and determinism. I think, in fact, that the above argument for the incompatibility of free will and determinism can be turned into a valid argument by a minor modification of Rule β.[8] Suppose that, instead of defining 'N*p*' as '*p* and no one has, or ever had, any choice about *p*'—that is, as '*p* and every region to which anyone has, or ever had, access overlaps *p*'—, we were to define 'N*p*' as follows:

> '*p* and every region to which anyone has, or ever had, *exact* access *is a sub-region of p*'.

One has *exact* access to a region if one has access to it *and to none of its proper subregions*. Intuitively, one has exact access to *p* if one can ensure the truth of *p* but of nothing "more definite." The properties of the "exact access" relation differ from those of the "access" relation in several important ways. If I am inside a region, I do not in general have exact access to that region. (This is an understatement: the only region I am inside and have exact access to is the actual world.[9]) If I have exact access to a region, then, by definition, I have exact access to none of its (proper) superregions.[10] If I have exact access to the region of logical space in which Hillary Clinton proves Goldbach's Conjecture,

it follows that I do *not* have exact access to the region in which *someone* proves Goldbach's Conjecture—although it follows that I *do* have *access* to that region. It is, unfortunately, impossible to give a plausible example of a non-actual region to which I have exact access. Suppose that, although I do not throw the dart, it is within my power to ensure that it hit the board—*and* that, for no proper part of the board is it within my power to ensure that the dart hit that part. Do I have exact access to a region in which the dart hits the board? Presumably not, for presumably I have access to a region in which the dart hits the board *and* I exclaim, "Ah!" For one to have exact access to the non-actual region *p* it must be the case that one can ensure the actuality of *p* but not the joint actuality of *p* and any logically independent region. If one could ensure the actuality of some non-actual *world*, one would have exact access to that world, of course, but obviously no one can do that—or no one but God. Still, it seems evident that there must be regions of logical space to which any given human being has exact access, simply because a human being's ability to ensure the truth of things, to "fine tune" his actions and their consequences, must come to an end somewhere.

Consider now our operator 'N', redefined as I have suggested. I think that this is what I was trying to capture when I defined 'N*p*' as '*p* and no one has, or ever had, any choice about *p*'. What McKay and Johnson's counterexample shows is that the concept 'not having a choice about' has the wrong logical properties to capture the idea I wanted to capture—the idea of the *sheer inescapability* of a state of affairs. But if 'N' is redefined in the way I have proposed, the redefined 'N' does capture this idea. If every region to which I have access overlaps *p*, it may nevertheless be true that there is some action I can perform such that, if I did, then *p might* be false. But if every region to which I have *exact* access *is a subregion of p*, every action I can perform is such that, if I did perform it, *p* would be true: it is not the case that *p* might be false.

Now if 'N' is re-defined as I have suggested, Rule β is valid—for the simple reason that every set that is a subset of both *p* and $p \supset q$ (that is, of $p \,\&\, q$) is a subset of *q*. Thus, if every region of logical space to which anyone has exact access is within both *p* and $p \supset q$, every region of logical space to which anyone has exact access is within *q*. (And, of course, Rule α is valid: every region of logical space to which anyone has exact access is a region of logical space.)

What about the two premises of the argument for the incompatibility of free will and determinism? These both seem true—or at least the reasons for thinking them true are no worse than they were on the "no choice" understanding of 'N'. Every region of logical space to which anyone has exact access will be a subregion of P_0; every region of logical space to which anyone has exact access will be a subregion of L. (The compatibilist will disagree. The compatibilist will define 'is able' in some way—will no doubt employ some version of the "conditional analysis of ability"—that will have the consequence that each of us is *able* to perform various acts, such that, if he or she did perform them,

then the conjunction of P_0 and L would be false. Thus, the compatibilist will argue, we do have exact access to regions that are not subregions of both P_0 and L. But this is an old dispute, and I have nothing new to say about it. I will say only this—and this is nothing new. The compatibilist's "move" is contrived and ad hoc; it is "engineered" to achieve the compatibility of free will and determinism; it *seems* that our freedom can only be the freedom to add to the actual past[11]; it *seems* that our freedom can only be the freedom to act in accordance with the laws of nature.)

It seems, therefore, that I now have what I thought I had when I thought Rule β was valid on the "no choice" understanding of 'N': a valid argument for the incompatibility of free will and determinism whose premises seem to be true. And this, *mutatis mutandis*, is all that can be asked of any philosophical argument. At any rate, no more can be said for any known philosophical argument than this: it is valid and its premises seem to be true.

II

Free will, then, seems to be incompatible with determinism. But, as many philosophers have noted, it also seems to be incompatible with *in*determinism. The standard argument for this conclusion (which I have called the *Mind* Argument because it has appeared so frequently in the pages of *Mind*) goes something like this:

> If indeterminism is to be relevant to the question whether a given agent has free will, it must be because the acts of that agent cannot be free unless they (or perhaps their immediate causal antecedents) are undetermined. But if an agent's acts are undetermined, then *how* the agent acts on a given occasion is a matter of chance. And if how an agent acts is a matter of chance, the agent can hardly be said to have free will. If, on some occasion, I had to decide whether to lie or to tell the truth, and if, after much painful deliberation, I lied, my lie could hardly have been an act of free will if whether I lied or told the truth was a matter of chance. To choose to lie rather than tell the truth is a *free* choice only if, immediately before the choice was made, it was up to the agent whether he lied or told the truth. That is to say, before the choice was made, the agent must have been able to lie and able to tell the truth. And if an agent is faced with a choice between lying and telling the truth, and if it is a *mere matter of chance* which of these things the agent does, then it cannot be up to the agent which of them he does.

(At any rate, this is one way to formulate the *Mind* Argument. Other statements of the argument are available, including some that do not appeal to the concept of chance. I will presently return to this point.) In *An Essay on Free Will*, I tried to show that the *Mind* Argument depended on the "unrevised" version of

Rule β. If this is correct, then, since "unrevised β" is invalid, the *Mind* Argument is invalid. But perhaps I was wrong to think that the *Mind* Argument depended on "unrevised β," at least in any essential way. Perhaps the *Mind* Argument depends only on the employment of some rule of inference *of the same general sort* as "unrevised β." Perhaps, indeed, the Mind Argument could be re-written so as to depend only on "revised β." I will not consider these possibilities. I will not try to answer the question whether the *Mind* Argument is in fact valid. I have a different project. I wish to consider the *Mind* Argument in a very informal, intuitive form, to contend that in this intuitive form the argument has a great deal of plausibility, and to use this contention as the basis of an argument for the conclusion that the concept of *agent causation* is entirely irrelevant to the problem of free will. This is no trivial conclusion. Most philosophers who have thought carefully about the problem of free will maintain that the concept of agent causation is incoherent—and perhaps also maintain that if, *per impossibile*, this concept were coherent, it would be contrary to naturalism or to some other important philosophical commitment to suppose that it applied to anything in the real world. A sizable and respectable minority of the philosophers who have thought carefully about the problem of free will maintain that the concept of agent causation is coherent and, moreover, that agent causation is real and figures in an essential way in the acts of free agents. But almost everyone seems to think that if there really *were* such a thing as agent causation, its reality would constitute a solution to the problem of free will. I am going to try to show that even if agent causation exists, even if it is an element in the acts of free agents, the problem of free will is just as puzzling as it would have been if no one had ever thought of the idea of agent causation. I am going to try to show that even if agent causation is a coherent concept and a real phenomenon, and we know this, this piece of knowledge will be of no help to the philosopher who is trying to decide what to say about free will.

I begin my argument by characterizing the problem of free will and the concept of agent causation.

The problem of free will in its broadest outlines is this. Free will seems to be incompatible both with determinism and indeterminism. Free will seems, therefore, to be impossible. But free will also seems to exist. The impossible therefore seems to exist. A solution to the problem of free will would be a way to resolve this apparent contradiction. There would seem to be three forms a solution could take, three ways in which one might try to resolve the apparent contradiction. One might try to show, as the compatibilists do, that—despite appearances—free will is compatible with determinism. Or one might try to show, as many incompatibilists do, that—despite appearances—free will is compatible with indeterminism. Or one might try to show, as many "hard determinists" do, that the apparent reality of free will is mere appearance. (To be reasonably plausible, a solution of the third type would probably have to incorporate some sort of argument for the conclusion that moral responsibility does not, as it appears to, require free will—or else an argument for the conclusion

that a belief in the reality of moral responsibility is not, as it appears to be, an indispensable component of our moral and legal and political thought.) This is the problem to which, in my view, agent causation is irrelevant. (Perhaps there is some *other* problem that could reasonably be called 'the problem of free will' and to which agent causation *is* relevant. I can only say that if there is such a problem, I don't know what it is.)

Agent causation is, or is supposed to be, a relation that agents—thinking or rational *substances*—bear to events. Agent causation is opposed to *event* causation, a relation that events bear to events. The friends of agent causation hold that the causes of some events are not (or are only partially) earlier events. They are rather substances—not *changes* in substances, which are of course events, but "the substances themselves." Thus, they say, Thomas Reid caused the movements of his fingers when he wrote the sentence, "There is no greater impediment to the advancement of knowledge than the ambiguity of words." These movements, they insist, were caused simply by *Reid*, and not by any change in Reid. Or, speaking more carefully, since they are aware on empirical grounds that these movements were in fact caused by changes in Reid's hand and arm and spinal cord and brain, they will say that there were *some* events, events that occurred no more than a few seconds before these movements and were among their causal antecedents, events that presumably occurred within the motor centers of Reid's brain, that were caused by Reid and not by any prior events. Speaking even more carefully, they may say that at any rate there were causal antecedents of the movements of Reid's fingers to whose occurrence Reid, Reid himself, the thinking substance, *contributed causally*—thus allowing the possibility that earlier events in Reid's brain *also* contributed causally to the occurrence of these events.

Let this suffice for a characterization of the problem of free will and the concept of agent causation. Now how is the concept supposed to figure in a solution of the problem? In this wise, I believe: the reality of agent causation is supposed to entail that free will and indeterminism are compatible. The idea is something like this. A certain event happens in Reid's brain, an event that, through various intermediate causes, eventually produces a bodily movement that constitutes some voluntary action of Reid's—say, his writing the sentence, "There is no greater impediment to the advancement of knowledge than the ambiguity of words." (Perhaps we need not attempt to explain the notion of a bodily movement's "constituting" a voluntary action. The idea is illustrated by this example: certain movements of Reid's arm and hand and fingers constitute his writing the sentence, "There is no greater impediment, *etc*.") And Reid is, let us suppose, the agent-cause of the aforementioned brain-event that was a causal antecedent of his writing this sentence—or at any rate he contributes agent-causally to its occurrence. (From this point on, I will neglect the distinction between agent-causing an event and contributing agent-causally to its occurrence.) The action, or the event that is Reid's performing it, is not determined by the state of the universe at any time before the antecedent brain-event oc-

curred. (And why not? Well, because the event that was his agent-causing the antecedent brain-event was not determined to occur by any prior state of the universe. And if that event—his agent-causing the antecedent brain-event—had not occurred, his hand and fingers would not have moved and he would not have written the sentence.) And yet it is as obviously true as anything could be that he is responsible for this event, for he was its cause: it occurred because *he* caused it to occur. It was therefore an act of free will, and free will is therefore consistent with indeterminism.

In the sequel, I will take it for granted that the relevance of the concept of agent causation to the problem of free will is supposed to be found in the supposed fact that the reality of agent causation entails that free will is compatible with indeterminism. And I will take it for granted that the argument of the preceding paragraph is a fair representation of the argument that is supposed to establish this compatibility. If there is some other reason agent causation is supposed to be relevant to the problem of free will, or if the argument of the preceding paragraph is a poor or incomplete representation of the reasons for supposing that the concept of agent causation can be used to establish the compatibility of free will and indeterminism, then the argument of the remainder of this essay will be at best incomplete and at worst entirely beside the point.

In my view, this argument does not succeed in showing that the reality of agent causation entails the compatibility of free will and indeterminism. Its weak point, I believe, is the reasoning contained in its last two sentences: "And yet it is as obviously true as anything could be that [Reid] is responsible for [the antecedent brain-event], for he was its cause: it occurred because *he* caused it to occur. It was therefore an act of free will, and free will is therefore consistent with indeterminism." It is not my plan to make anything of the fact that Reid knew even less than I about what goes on in the motor centers of human brains—or of the fact that other agents, agents who act freely if anyone does, do not even know that they *have* brains. Any doubts about the argument that might be based on these facts have to my mind been adequately answered by Chisholm, and I shall not bother about them.[12] Nor shall I raise questions about the cause of the event "its coming to pass that Reid is the agent-cause of the antecedent brain-event."[13] Again, I think Chisholm has seen what the friends of agent causation should say about the cause of this event, to wit, that Reid was its agent-cause—and was, moreover, the agent-cause of the event "its coming to pass that Reid is the agent-cause of the event 'its coming to pass that Reid is the agent-cause of the antecedent brain-event'," and so *ad infinitum*.[14] Some may object to the thesis that, as an indispensable component of his writing a certain sentence, Reid, without being aware of it, became the agent cause of an infinite number of events; I don't.

In order to see what I *do* object to in the argument, let us return to the question why some have thought that free will was incompatible with indeterminism. Let us, that is, return to the "mere matter of chance" argument. Let us try to state this argument more carefully. (In *An Essay on Free Will*, I had a

very short way with any attempt to state the *Mind* argument in terms of an undetermined act's being a random or chance occurrence.[15] I argued there that the words 'random' and 'chance' most naturally applied to *patterns* or *sequences* of events, and that it was therefore not clear what these words could mean if they were applied to single events. It will be evident from what follows that I no longer regard this argument as having any merit.) Let us suppose undetermined free acts occur. Suppose, for example, that in some difficult situation Alice was faced with a choice between lying and telling the truth and that she freely chose to tell the truth—or, what is the same thing, she seriously considered telling the truth, seriously considering lying, told the truth, and was able to tell the lie she had been contemplating. And let us assume that free will is incompatible with determinism, and that Alice's telling the truth, being a free act, was therefore undetermined. Now suppose that immediately after Alice told the truth, God caused the universe to revert to precisely its state one minute before Alice told the truth (let us call the first moment the universe was in this state 't_1' and the second moment the universe was in this state 't_2'), and then let things "go forward again." What would have happened the second time? What would have happened after t_2? Would she have lied or would she have told the truth? Since Alice's "original" decision, her decision to tell the truth, was undetermined—since it was undetermined whether she would lie or tell the truth—, her "second" decision would also be undetermined, and this question can therefore have no answer; or it can have no answer but, "Well, although she would either have told the truth or lied, it's not the case that she would have told the truth and it's not the case that she would have lied; lying is not what she would have done, and telling the truth is not what she would have done. One can say only that she *might* have lied and she *might* have told the truth."

Now let us suppose that God *a thousand times* caused the universe to revert to exactly the state it was in at t_1 (and let us suppose that we are somehow suitably placed, metaphysically speaking, to observe the whole sequence of "replays"). What would have happened? What should we expect to observe? Well, again, we can't say what would have happened, but we can say what would *probably* have happened: sometimes Alice would have lied and sometimes she would have told the truth. As the number of "replays" increases, we observers shall—almost certainly—observe the ratio of the outcome "truth" to the outcome "lie" settling down to, converging on, some value.[16] We may, for example, observe that, after a fairly large number of replays, Alice lies in thirty percent of the replays and tells the truth in seventy percent of them—and that the figures 'thirty percent' and 'seventy percent' become more and more accurate as the number of replays increases. But let us imagine the simplest case: we observe that Alice tells the truth in about half the replays and lies in about half the replays. If, after one hundred replays, Alice has told the truth fifty-three times and has lied forty-eight times,[17] we'd begin strongly to suspect that the

figures after a thousand replays would look something like this: Alice has told the truth four hundred and ninety-three times and has lied five hundred and eight times. Let us suppose that these are indeed the figures after a thousand replays. Is it not true that as we watch the number of replays increase, we shall become convinced that what will happen in the *next* replay is a matter of chance? (The compulsive gamblers among us might find themselves offering bets about what Alice would do in the next replay.) If we have watched seven hundred and twenty-six replays, we shall be faced with the inescapable impression that what happens in the seven-hundred-and-twenty-seventh replay will be due simply to chance. Is there any reason we should resist this impression? Well, we certainly know that there is nothing we could learn about the situation that could undermine the impression, for we already know everything that is relevant to evaluating it: we know that the outcome of the seven-hundred-and-twenty-seventh replay will not be determined by its initial state (the common initial state of all the replays) and the laws of nature. Each time God places the universe in this state, both "truth" and "lie" are consistent with the universe's being in this state and the laws of nature. A sheaf of possible futures (possible in the sense of being consistent with the laws) leads "away" from this state, and, if the sheaf is assigned a measure of 1, surely, we must assign a measure of 0.5 to the largest sub-sheaf in all of whose members Alice tells the truth and the same measure to the largest sub-sheaf in all of whose members she lies. We must make this assignment because it is the only reasonable explanation of the observed approximate equality of the "truth" and "lie" outcomes in the series of replays. And if we accept this general conclusion, what other conclusion can we accept about the seven-hundred-and-twenty-seventh replay (which is about to commence) than this: each of the two possible outcomes of this replay has an objective, "ground-floor" probability of 0.5—and there's nothing more to be said? And this, surely, means that, in the strictest sense imaginable, the outcome of the replay will be a matter of chance.

Now, obviously, what holds for the seven-hundred-and-twenty-seventh replay holds for all of them, including the one that wasn't strictly a *re*play, the initial sequence of events. But this result concerning the "initial replay", the "play," so to speak, should hold whether or not God bothers to produce any replays. And if He does not—well, that's just the actual situation. Therefore, an undetermined action is simply a matter of chance: if it was undetermined in the one, actual case whether Alice lied or told the truth, it was a mere matter of chance whether she lied or told the truth. If we knew beforehand that the objective, "ground-floor" probabilities of Alice's telling the truth and Alice's lying were both 0.5, then (supposing our welfare depended on her telling the truth) we could only regard ourselves as *fortunate* when, in the event, she told the truth. But then how can we say that Alice's telling the truth was a free act? If she was faced with telling the truth and lying, and it was a mere matter of chance which of these things she did, how can we say that—and this is essen-

tial to the act's being free—she was *able* to tell the truth and *able* to lie? How could anyone be able to determine the outcome of a process whose outcome is a matter of objective, ground-floor chance?

This is the plausible, intuitive version of the *Mind* Argument that I have promised to discuss. What I must do now is show that the concept of agent causation cannot be used to undermine the intuitive plausibility of this argument.

Let us suppose that when Alice told the truth, she agent-caused certain brain-events that, in due course, resulted in those movements of her lips and tongue that constituted her telling the truth. And let us again suppose that God has caused the universe to revert to precisely its state at t_1, and that this time Alice has lied. I do not see how to avoid supposing that in this "first replay," Alice *freely* lied—for if one has to choose between telling the truth and lying, and if one freely chooses to tell the truth, then it must be the case that if one had chosen instead to lie, the choice to lie would have been a free act. (One cannot say that an agent faces exactly two continuations of the present, in one of which he tells the truth but was able to lie and in the other of which he lies and was *un*able to tell the truth.) Now if Alice's lie in the first replay was a free act, she must—according to the friends of agent causation—have been the agent-cause of some among the causal antecedents of the bodily movements that constituted her lying. And so, of course, it will be, *mutatis mutandis*, in each successive replay. If God produces one thousand replays, and if (as I have tacitly been assuming) the state of the universe at t_1—the common initial state of all the replays—determines that Alice will *either* tell the truth or lie, then, in each replay, Alice will *either* agent-cause cerebral events that, a second or so later, will result in bodily movements that constitute her telling the truth or agent-cause cerebral events that, a second or so later, will result in bodily movements that constitute her lying. She will, perhaps, agent-cause events of the "truth antecedent" sort five hundred and eight times and events of the "lie antecedent" sort four hundred and ninety-three times. Let us suppose once more that we are somehow in a position to observe the sequence of replays. We may again ask the question, "Is it not true that as we watch the number of replays increase, we shall become convinced that what will happen in the *next* replay is a matter of chance?" I do not see why we should not become convinced of this. And what might we learn, what is *there* for us to learn, that should undermine this conviction? What should lead us to say that the outcome of the next replay, the seven hundred and twenty-seventh, will not be a matter of chance? What should lead us to say that it is anything other than a matter of chance whether Alice will agent-cause truth-antecedent cerebral events or lie-antecedent cerebral events in the about-to-occur seven-hundred-and-twenty-seventh replay? Well, one might say this: If it turns out that Alice agent-causes truth-antecedent cerebral events, this will not be a matter of chance because it will be she, *Alice*, who is the cause of the event "its coming to pass that Alice agent-causes truth-antecedent cerebral events." But have we not got every reason to regard the occurrence of *this* event—that is, the occurrence of "its coming to pass that

Alice agent-causes the event 'its coming to pass that Alice agent-causes truth-antecedent cerebral events"'—as a matter of chance? If the three events "the truth-antecedent cerebral events"/ "its coming to pass that Alice agent-causes the truth-antecedent cerebral events"/ "its coming to pass that Alice agent-causes the event 'its coming to pass that Alice agent-causes truth-antecedent cerebral events"' are the first three terms of an infinite series of agent-caused events, is not the simultaneous occurrence of all the events in this sequence (as opposed to the simultaneous occurrence of all the events in an infinite sequence of agent-caused events whose first member is "lie-antecedent cerebral events") a mere matter of chance?

Nothing we could possibly learn, nothing God knows, it would seem, should lead us to distrust our initial inclination to say that the outcome of the next replay will be a matter of chance. If this much is granted, the argument proceeds as before, in serene indifference to the fact that we are now supposing Alice to be the agent-cause of various sets of cerebral events that are antecedents of the bodily movements that constitute her acts. And the argument proceeds to this conclusion: if it is undetermined whether Alice will tell the truth or lie, then—*whether or not* Alice's acts are the results of agent-causation—it is a mere matter of chance whether she will tell the truth or lie. And if it is a mere matter of chance whether she will tell the truth or lie, where is Alice's free will with respect to telling the truth and lying? If one confronts a choice between A and B and it is a matter of chance whether one will choose A or B, how can it be that one is *able* to choose A?

I close with an example designed to convince you of this.

You are a candidate for public office, and I, your best friend, know some discreditable fact about your past that, if made public, would—and should—cost you the election. I am pulled two ways, one way by the claims of citizenship and the other by the claims of friendship. You know about my situation and beg me not to "tell." I know (perhaps God has told me this) that there exist exactly two possible continuations of the present—the actual present, which includes your begging me not to tell and the emotional effect your appeal has had on me—, in one of which I tell all to the press and in the other of which I keep silent; and I know that the objective, "ground-floor" probability of my "telling" is 0.43 and that the objective, "ground-floor" probability of my keeping silent is 0.57. Am I in a position to promise you that I will keep silent?—knowing, as I do, that if there were a million perfect duplicates of me, each placed in a perfect duplicate of my present situation, forty-three percent of them would tell all and fifty-seven percent of them would hold their tongues? I do not see how, in good conscience, I could make this promise. I do not see how I could be in a position to make it. But if I believe that I am able to keep silent, I should, it would seem, regard myself as being in a position to make this promise. What more do I need to regard myself as being in a position to promise to do X than a belief that I am *able* to do X? Therefore, in this situation, I should not regard myself as being able to keep silent. (And I cannot see on what grounds

third-person observers of my situation could dispute this first-person judgment.) Now suppose God vouchsafes me a further revelation: "Whichever thing you do, whether you go to the press or keep silent, you will be the agent-cause of events in your brain that will result in the bodily movements that constitute your act." Why should this revelation lead me to conclude that I am in a position to promise to keep silent—and therefore that I am able to keep silent? Its content simply doesn't seem to be relevant to the above argument for the conclusion that it is false that I am able to keep silent. I confess I believe there *is* something wrong with this argument. (I expect I believe this because I fervently *hope* that there is something wrong with it.) But it seems clear to me that if there is, as I hope and believe, something wrong with the argument, its flaw is not that it overlooks the possibility that my actions have their root in agent-causation.[18]

Notes

1. See Peter van Inwagen, *An Essay on Free Will* (Oxford: at the Clarendon Press, 1983), pp. 93–104.
2. Or this will do as a first approximation to the truth. But the statement in the text is not literally true, since at least one of the two premises is a contingent truth. ('P_0' is a contingent truth, and 'NP_0', which has 'P_0' as a conjunct, is therefore a contingent truth. 'L' is *probably* a contingent truth, and 'NL' is therefore probably a contingent truth.) Here is a more careful statement. If the two rules of inference are valid, then an argument identical in appearance with the argument in the text can be constructed in any possible world and premises (4) and (6) of any of these arguments will be true in the possible world in which it is constructed if 'P_0' expresses a proposition that describes the state of the world (= 'universe') in that possible world before there were any human beings, and 'L' expresses the proposition that is the conjunction of all propositions that are laws of nature in that possible world. Thus it can be shown (if the two rules of inference are valid) with respect to each possible world that if determinism is true in that world, then none of its inhabitants has any choice about anything. And if this can be shown with respect to each possible world, then free will is incompatible with determinism.
3. That is, if a region of logical space occupies 23.37 percent of the whole of logical space, the probability of its being actual (containing the actual world) is 0.2337: the "intrinsic probability" of a proposition that is true in just that region of logical space is 0.2337. See my "Probability and Evil," in *The Possibility of Resurrection and Other Essays in Christian Apologetics* (Boulder: Westview Press, 1997), pp. 69–87.
4. In this paper, the symbols '*p*' and '*q*' and so on will sometimes be schematic letters representing sentences and sometimes variables ranging over propositions or regions of logical space. Although I normally deprecate this sort of logical sloppiness, it does have its stylistic advantages, and it is easily eliminable at the cost of a little verbal clutter. Similar remarks apply to '&' and '⊃'.
5. Thomas McKay and David Johnson, "A Reconsideration of an Argument against Compatibilism," *Philosophical Topics* 24 (1996), pp. 113–122.

6. The McKay/Johnson counterexample to Agglomeration is not a counterexample to β—although, since, since the validity of β entails the validity of Agglomeration, the existence of a counterexample to Agglomeration entails the existence of counterexamples to β.
7. It implies more than this. It implies something about knowledge, generally knowledge of cause and effect. If *p* is true, and if *p* would be false if I did X (which I was able to do), for me to have a choice about the truth-value of *p*, I must have known (or at least be such that I *should* have known) that doing X would result in the falsity of *p*.
8. Other ways to repair the argument have been suggested. One of these ways—it is similar to my own proposal—has been suggested by McKay and Johnson themselves. See "A Reconsideration of an Argument against Compatibilism," pp. 118–121. For a different suggestion, see Alicia Finch and Ted A. Warfield, "The *Mind* Argument and Libertarianism," *Mind* 107 (1998), pp. 515–528.
9. This statement assumes that no non-actual world is as close to the actual world as the actual world is to itself. Without this assumption, we should have to say: the only region I am inside and have exact access to is the set of worlds that are as close to the actual world as it is to itself.
10. Suppose I have exact access to *r*. Then I have access to *r*. Let R be any (proper) superregion of *r*. If I have exact access to R, I have exact access to a region and to one of its proper subregions (*r*)—which is contrary to the definition of exact access.
11. *Cf.* Carl Ginet, *On Action* (Cambridge, Cambridge University Press, 1990), pp. 102–103.
12. "Freedom and Action," in Keith Lehrer, ed., *Freedom and Determinism* (New York: Random House, 1966), pp. 11–44. See pp. 20–21.
13. The event "its coming to pass that Reid is the agent-cause of the antecedent brain-event" is the same event as "Reid's acquiring the property *being the agent-cause of the antecedent brain event*." Presumably, there is a moment of time before which Reid has not agent-caused the antecedent brain-event and after which he has, and that is the moment at which this event occurs.
14. At any rate, I *believe* that Chisholm has considered this problem and has defended the "and so *ad infinitum*" solution. But I have been unable to find this solution in his writings.
15. Pp. 128–29.
16. "Almost certainly" because it is *possible* that the ratio not converge. Possible but most unlikely: as the number of replays increases, the probability of "no convergence" tends to 0.
17. After one hundred replays, Alice has told the truth or lied one hundred and one times.
18. I am grateful to Ted A. Warfield (*il miglior fabbro*) for reading Part I of this paper and for offering valuable criticisms. I hope I have made good use of them.