Midwest Studies in Philosophy, XXVIII (2004)

Freedom to Break the Laws

PETER VAN INWAGEN

Philosophers are unable to agree about free will. Some are determinists[1] who deny free will, some determinists who affirm free will. Some philosophers think that free will is incompatible with determinism *and* with indeterminism—and hence that free will is impossible—while others say that we are free and that our free actions are and must be undetermined; yet others say that we are free and that our free actions are and must be determined. Some philosophers believe that acts of free will involve a special kind of causation, agent causation (whereby a substance causes alterations in the world without itself undergoing any alteration), and others respond to appeals to agent causation with incredulous stares. Some say that free will is an unintelligible notion, and others say that, whether the *thing* free will exists or not, the *concept* "free will" is a paradigm of intelligibility. Some say that although free will is incompatible with determinism, this fact is of little consequence because moral responsibility (which is what is really at issue in debates about free will) *is* compatible with determinism; their opponents reply that it is *evident* that moral responsibility cannot exist without free will. I could go on, but I trust I have made my point: the problem of free will is a typical philosophical problem.

1

Disagreement in philosophy is pervasive and irresoluble. There is almost no thesis in philosophy about which philosophers agree. If there is any philosophical thesis

1. Or "near-determinists": with a polite bow in the direction of quantum mechanics, these philosophers insist that there is no more indeterminism in the workings of a human being than there is in the workings of a digital computer.

that all or most philosophers affirm, it is a negative thesis: that formalism is not the right philosophy of mathematics, for example, or that knowledge is not (simply) justified, true belief.[2] (See the long quotation from David Lewis a few paragraphs on: "Gödel and Gettier may have done it.")

That is not how things are in the physical sciences. I concede that the "cutting edge" of elementary-particle physics looks a lot like philosophy in point of pervasive and fundamental disagreement among its respected practitioners. But there is in physics a large body of settled, usable, uncontroversial theory and of measurements known to be accurate within limits that have been specified. The cutting edge of philosophy, however, is pretty much the whole of it.[3]

That is not how things are in history. I concede that the historian Peter Geyl was making a true and important point when he said that history was argument without end. Nevertheless, there is no controversy whatever about whether Queen Anne is dead. Unlike the physical sciences, history does not have "a large body of settled, usable, uncontroversial theory" at its disposal but, like the physical sciences, it does have a large body of established and incontrovertible fact to work with. In philosophy, however, there is neither settled theory nor incontrovertible fact.[4]

And that is not how things are in mathematics, the biological sciences, economics, linguistics, archaeology. . . . The fundamental, pervasive, and irresoluble disagreement that afflicts philosophy is certainly uncommon. It is a defensible position that it is unique.

How do philosophers react when this state of affairs comes to their attention? In the seventeenth and eighteenth centuries, generally by proclaiming some new philosophical method that would finally put the feet of philosophy on the sure path of science. Such proclamations have been rare or non-existent for quite a long time now, and understandably so. Present-day analytical philosophers tend simply not to permit the fact that philosophical disagreement is irresoluble to come to their attention. One analytical philosopher who did not ignore this fact was David

2. I used to think that all philosophers, or at least all philosophers of mathematics and all philosophically minded logicians, accepted Church's Thesis (which is a philosophical thesis). I was wrong. See László Kalmár, "An Argument against the Plausibility of Church's Thesis," in A. Heyting (ed.), *Constructivity in Mathematics* (Amsterdam: North Holland, 1959), pp. 72–80.

3. It is uncontroversial that the general theory of relativity may be used to calculate the loss of energy due to gravitational radiation in a pair of rotating neutron stars: that that theory, properly applied, will yield accurate results in the "regime" to which such a system belongs. It is uncontroversial that the statement that the rest-mass of the electron is 9.11×10^{-31} kg is accurate to within the displayed number of decimal places.

4. For a remark on one attempt of philosophers to find themselves a body of uncontroversial data, see the quotation from Lewis that we are moving toward in the text (the remark on "linguistic intuition"). Another such attempt is represented by Husserl's slogan, "back to the phenomena." *Every* such attempt has generated the following reaction: some philosophers deny that the supposed foundational data are data; others deny that they are foundational; still others say that even if they are data and are foundational (in the sense that they are self-evident and require no justification beyond themselves), not much philosophy can be based on them.

Lewis.[5] Here is the long quotation I have been promising. It is from the Introduction to the first volume of his collected *Philosophical Papers*.[6]

> The reader in search of knock-down arguments in favor of my theories will go away disappointed. Whether or not it would be nice to knock disagreeing philosophers down by sheer force of argument, it cannot be done. Philosophical theories are never refuted conclusively. (Or hardly ever. Gödel and Gettier may have done it.) The theory survives its refutation—at a price. Argle has said what we accomplish in philosophical argument: we measure the price. Perhaps that is something we can settle more or less conclusively. But when all is said and done, and all the tricky arguments and distinctions and counterexamples have been discovered, presumably we will still face the question which prices are worth paying, which theories are on balance credible, which are the unacceptably counterintuitive consequences and which are the acceptably counterintuitive ones. On this question we may still differ. And if all is indeed said and done, there will be no hope of discovering still further arguments to settle our differences.
>
> It might be otherwise if, as some philosophers seem to think, we had a sharp line between "linguistic intuition," which must be taken as unchallengeable evidence, and philosophical theory, which must at all costs fit this evidence. If that were so, conclusive refutations would be dismayingly abundant. But, whatever may be said for foundationalism in other subjects, this foundationalist theory of philosophical knowledge seems ill-founded in the extreme. Our "intuitions" are simply opinions; our philosophical theories are the same. Some are commonsensical, some are sophisticated; some are particular, some general; some are more firmly held, some less. But they are all opinions, and a reasonable goal for a philosopher is to bring them into equilibrium. Our common task is to find out what equilibria there are that can withstand examination, but it remains for each of us to come to rest at one or another of them. If we lose our moorings in everyday common sense, our fault is not that we ignore part of our evidence. Rather, the trouble is that we settle for a very inadequate equilibrium. If our official theories disagree with what we cannot help thinking outside the philosophy room, then no real equilibrium has been reached. Unless we are doubleplusgood doublethinkers, it will not last. And it should not last, for it is safe to say that in such a case we will believe a great deal that is false.
>
> Once the menu of well-worked-out theories is before us, philosophy is a matter of opinion. Is that to say that there is no truth to be had? Or that the truth is of our own making, and different ones of us can make it differently? Not at all! If you say flatly that there is no god, and I say that there are countless gods but none of them are our worldmates, then it may be that neither

5. Another is Colin McGinn. See his *Problems in Philosophy: The Limits of Inquiry* (Oxford: Blackwell, 1993).

6. David Lewis, *Philosophical Papers, Vol. I* (New York and Oxford: Oxford University Press, 1983), pp. x–xi. There are several footnotes to this text. The reader may wish to consult the original.

> of us is making any mistake of method. We may each be bringing our opinions to equilibrium in the most careful possible way, taking account of all the arguments, distinctions, and counterexamples. But one of us, at least, is making a mistake of fact. Which one is wrong depends on what there is.

Let me say something to tie what I have been talking about (pervasive and irresoluble philosophical disagreement) to what Lewis was talking about (the absence of knock-down arguments from philosophy).[7]

What could put an end to disagreement in philosophy if not knock-down arguments? Philosophical agreement will come to pass when, and only when, for each important philosophical thesis, there is a knock-down argument either for that thesis or for its denial.[8] In saying this, I suppose that philosophical theses are, at least for the most part, genuine propositions (proper objects of affirmation or denial; possessed of truth-values).[9] (I suppose this, as does Lewis: "mistake of fact"; "depends on what there is.") If they are pseudo-propositions of some sort, then the resolution of philosophical disagreement (or pseudo-disagreement) will require not argument but therapy.[10]

The question whether philosophical theses are genuine propositions is important enough in relation to the topic of philosophical disagreement to be worth a brief digression.

What can be said in support of the philosophical thesis that philosophical theses are genuine propositions? I offer the following argument. At least *some* philosophical theses must be genuine propositions. For consider the proposition that all philosophical theses are pseudo-propositions. This proposition is itself a philosophical thesis, for philosophy is a part of its own subject-matter: "What is a philosophical thesis?" is a philosophical question. This proposition is therefore a pseudo-proposition if it is true. The best course for those who *want* to say that all philosophical theses are pseudo-propositions (and who have seen that if they do say what they want to say, they will be either affirming a falsehood or attempting to affirm a pseudo-proposition) is to affirm some instance of the following schema: The members of a certain proper subset ϕ of the set of philosophical theses are pseudo-propositions; those philosophical theses that are *not* pseudo-propositions are theses *about* the members of ϕ, theses that ascribe certain intrinsic or

7. In the opening sentences of the quoted passage, Lewis speaks only of the (near) absence of knock-down *refutations* from philosophy. But, as the passage continues, it becomes clear that he means to assert the absence from philosophy of knock-down *arguments*—whether proofs or refutations.

8. The disjunction is exclusive. There cannot be a knock-down argument both for a thesis and for its denial. If, *per impossibile*, this situation did obtain, there would be, in Hume's phrase, "a mutual destruction of arguments," and neither argument would be knock-down after all.

9. I use the terms "proposition" and "thesis" as stylistic variants.

10. I use the term "pseudo-proposition" because of its important role in twentieth-century thinking about the nature of philosophy. But my use of the term should not be taken to imply that I suppose that there are things that appear to be propositions but aren't. I suppose, rather, that there are sentences that appear to express propositions but don't. The writers who first used the term "pseudo-proposition" used this term to designate what *I* should call "sentences that appear to express propositions but don't"—and used the word "proposition" to designate what I should call "sentences that express propositions."

relational properties to the members of ϕ ("Every member of ϕ is a pseudo-proposition" being of course one of them). Those who have reluctantly abandoned the thesis that all philosophical propositions are pseudo-propositions could, of course, try to find a revised version of this thesis that sounds more like the original than the revision *I* have suggested. They could insist upon applying the term "philosophy" only to the members of ϕ (whatever ϕ may be), and they could invent a new name for what otherwise would have been called "philosophical theses about the members of ϕ" (no doubt the name would be "metaphilosophical theses"). But this would be a merely verbal maneuver and would accomplish no more than any other merely verbal maneuver. Giving a new name to a certain class of philosophical theses (and denying them the old name "philosophical thesis") is not going to change the fact that, whatever they are called, all "metaphilosophical" theses are the objects of irresoluble disagreement. Philosophers exhibit no more tendency to agree about theses like "The sentence 'Human beings have free will' expresses no proposition" than they do about theses like "The sentence 'Human beings have free will' expresses a false proposition."[11]

Some philosophers have thought that they or their teachers had invented a new philosophical method, the application of which would (finally!) yield knock-down philosophical arguments. Others, most present-day analytical philosophers among them, eschew methodological questions and simply soldier on, applying the traditional methods of philosophy to "first-order" philosophical problems (that is, problems about things like the mind or morality or being and non-being, as opposed to problems concerning the nature of the philosophical enterprise). (They carefully define terms of art. They propose analyses of concepts. They advance counterexamples to analyses. They construct theories. They search out possible ambiguities in philosophically important words and phrases. They point out that this or that argument is not formally valid unless this or that proposition is added to its premises. They insist that one of the premises of an argument assumes the very point at issue. They contend that the philosophers who have favored a certain

11. Do I not by offering this argument contradict myself, at least pragmatically? Do I not represent the argument in the text (the argument whose conclusion is "At least some philosophical theses are genuine propositions") as a knock-down argument? No, I offer it only as an argument. I could write a "Wittgensteinian" reply, or, better, response, to it, the core of which would be something like this: "The philosopher who follows the proper method never asserts anything—not even the proposition that the philosopher who follows the proper method never asserts anything. All my assertions, even this one, are parlor tricks played with linguistic props, directed illusions, conceptual sleight-of-hand, whose purpose is to get my audience to see the supposed problems of philosophy in a new way. And this seeing-in-a-new-way is not a matter of belief. When people see things as I want them to see them, they will have gained no new beliefs and will have lost no old ones, just as people who now see a duck where a moment ago they saw a rabbit have gained no new beliefs and lost no old ones. I expect my audience, at the outset, to treat texts like the one you are now reading as comprising assertions; in the end, however, if I am successful and my audience does see things the way I want them to see them, they will no longer see anything I have said in the course of the cognitive therapy I have led them through as an assertion. Not that they will *have the belief* that the things I said were not assertions. . . ." (And so on. And so on.) And I do not claim that the argument presented in the text calls into question the validity (or whatever it should be called) of the line of thought (or whatever it should be called) represented by this response.

thesis would look much less favorably on one of its hitherto unnoticed consequences. They assign the burden of proof to one of the sides in a philosophical debate. They introduce into discussions of traditional philosophical questions considerations gleaned from the physical and biological sciences.)

And what *about* these present-day analytical philosophers, these philosophers who do not claim to present knock-down arguments as the result of having discovered some new philosophical method? Do they claim to present knock-down arguments as the result of having used traditional philosophical methods? Well, they rarely if ever make this claim in so many words. But I would point out that when they present the fruits of their researches in print, they employ in almost every paragraph of their books and essays phrases whose use suggests, and more than suggests, that their own philosophical work (they could hardly believe this about the central arguments of their opponents in philosophical debate) contains knock-down arguments ("I shall now show"; "This proof"; "The demonstration in the previous section"; "We see therefore"). It would seem that, like Kant and the logical positivists and all the other philosophers who have claimed to have discovered a new philosophical method, they do believe that there are knock-down arguments in philosophy. (And it is certainly true that they believe that there *could* be.) But whether one supposes that knock-down arguments in philosophy are the fruit of a new philosophical method or the fruit of some recent application of the perennial methods of philosophy (an application of these methods that one presumably supposes to be more painstaking and insightful than almost all the applications of these methods that are to be found in the history of philosophy), one must suppose that philosophical agreement will come to be only in the wake of the discovery of some knock-down philosophical arguments. One must, in fact, suppose this even if one believes that there can be no knock-down arguments in philosophy.

Lewis (in the quotation) and I (in the opening paragraphs of the present essay) are therefore talking about the same topic: the possibility of knock-down philosophical arguments and the possibility of agreement in philosophy are the same topic.[12]

As the quotation shows, Lewis is not a typical present-day analytical philosopher: he does not believe that knock-down philosophical arguments are possible, and he does not believe that it will ever happen that most philosophers agree that some given important and positive philosophical thesis is true. (It will "hardly ever" happen that most philosophers come to agree that a certain analysis or

12. My thesis about the relation between the possibility of knock-down philosophical arguments and the possibility of philosophical agreement is not based on any consideration peculiar to philosophy. I claim for it no plausibility beyond such plausibility as can be supplied by general reflection on the basis of agreement in any theoretical discipline. Consider those enviable theoretical disciplines in which "pervasive agreement" is the order of the day. Consider any proposition that, as the result of the researches of the experts in these disciplines, is generally agreed to be true ("The continents are in motion"; "The strands of the double helix are held together by hydrogen bonding" . . .). Will there not in every case be at least one knock-down argument for the truth of this proposition (or at least an argument that the experts *regard* as a knock-down argument)?

theory is wrong; it will never happen that most philosophers come to agree that a certain analysis or theory is right.)

What the philosopher can hope for, Lewis says, is to reach philosophical equilibrium. When one is in a state of philosophical equilibrium, one accepts certain answers to certain philosophical questions. One is, of course, aware of many philosophical "considerations"—arguments, definitions, principles, distinctions, and so on—that are (widely believed to be) relevant to the project of finding the correct answers to those philosophical questions to which one has accepted answers, and one believes that one has made a really serious attempt to survey all known philosophical considerations that are relevant to answering those questions. (One may or may not oneself have invented or discovered some of these considerations; one need not be an *original* philosopher to reach philosophical equilibrium.) And, finally, one is satisfied that one "knows what to say" about each of the considerations one is aware of that tells against or seems to tell against the answers one accepts. In particular, if the "consideration" is a valid argument for the denial of one of one's philosophical beliefs, one is prepared to say which of its premises one believes to be false and to explain why one believes them to be false. (An extreme case: "which of its premises" might be nothing more than the disjunction of its premises; but one would hope to do better than that.)

2

I see certain difficulties with Lewis's notion that a state of philosophical equilibrium is the best that one can hope for in philosophy. (The best one can hope for in respect of what? I mean something like this: the best that one can hope for in respect of warrant or justification or "positive epistemic status.") I will try to give a statement of the difficulties I think I see.

Let us call the set of philosophical propositions one accepts (at a given moment at which one is in a state of philosophical equilibrium) one's point of philosophical equilibrium (at that moment). Let us call two philosophers "co-workers" if their work is mutually relevant. (Whether two philosophers are co-workers will be a matter of degree, of course, and perhaps the degree to which two philosophers are co-workers will sometimes be a matter of opinion. But the notion of the degree to which two philosophers are co-workers is not an entirely subjective one. I would judge it to be uncontroversial that the degree to which Hartry Field and Martha Nussbaum are co-workers is negligible. Each is an excellent philosopher, but one would not expect the work of either to contain many citations of the work of the other.)

The following five theses seem to be true. (1) It is rare for two co-workers to reach the same point of equilibrium. (2) Pick some pair of co-workers at random (and to make our case as strong as possible, let's suppose that the two are anglophone analytical philosophers born in the same decade). Their points of philosophical equilibrium will probably be very different. (3) And not only very different, but inconsistent. And radically inconsistent: that is, these points of equilibrium could not be rendered consistent by "minor surgery," by making minor, superficial adjustments to either point of equilibrium or to both. (4) The intersec-

tion of all points of philosophical equilibrium will be a very small set of propositions, far too small to be itself anyone's point of philosophical equilibrium. (5) Consider, in fact, the intersection of the points of equilibrium that have been reached by ten randomly chosen anglophone analytical philosophers, born in the 1950s, who are all co-workers with one another to a fairly high degree. Even this intersection will (in all probability) be too small a set of propositions to be itself anyone's point of philosophical equilibrium.

Suppose I take a few moments to reflect seriously on the epistemological implications of these theses. Assume I have reached a certain point of philosophical equilibrium. What should occur to me in the matter of the degree of "warrant" (in the epistemological sense of the term) that this point of philosophical equilibrium enjoys (for me; in my present epistemic situation)? Let's assume that I know what this point of philosophical equilibrium I have reached is. Let's suppose, that is, that I am capable of setting out a certain list of philosophical propositions and that I am in a position to say truly: The point of philosophical equilibrium that I have reached contains the propositions in this list and all their logical consequences and no other propositions. (There may be difficulties with what I am asking you to suppose. How can I be certain that I accept all the logical consequences of some rich set of philosophical propositions? Might not this proposition about the membership of my point of philosophical equilibrium be itself a philosophical proposition, and, if so, might that fact not engender some paradox of self-reference? Let us ignore the possibility of such difficulties.) If I do know what my point of philosophical equilibrium is, then I think that it, that point of philosophical equilibrium, is *true* (that is, that all the beliefs it contains are true). This follows for the same reason as the reason for which it follows from my believing that snow is white and grass is green (and knowing that I believe these things) that I believe that the set of all the logical consequences of these two beliefs is true. (Maybe the valid deduction of this conclusion requires the additional premise that I understand the concept of the set of the logical consequences of a set of propositions. Consider it added.)

But what justifies me in accepting the proposition that my point of philosophical equilibrium is true? The totality of the arguments I endorse whose conclusions are members of my point of equilibrium? But consider: lots of other philosophers know about these arguments (they could state them as convincingly as I) and nevertheless occupy other points of philosophical equilibrium than mine. If the philosophical arguments I endorse have the power to confer warrant or justification (or whatever the most general terms of epistemic commendation are) on the proposition that my point of philosophical equilibrium is true, why don't these other philosophers come to rest at the same point of equilibrium as I? It would seem that no set of arguments could have the power to confer warrant on two inconsistent points of equilibrium. The arguments I endorse therefore confer warrant on one point of equilibrium (given that no other point of equilibrium is consistent with mine) or none. And if it is one and not none, I must suppose that it is mine. But why don't those of my co-workers who are familiar with and understand all the philosophical arguments I endorse *see* this? It is not easy to answer this question.

I should know how to answer it if I thought I was a significantly better philosopher than my co-workers whose points of equilibrium were inconsistent with mine (that is to say: if I thought I was a significantly better philosopher than *all* my co-workers).[13] I should know how to answer it if I thought that some fortunate combination of chance advantages had enabled me to "see" some complex of philosophically relevant factors (one so subtle that I have not yet succeeded in putting it into words and have thus been unable to communicate it to my co-workers) that confers warrant on my point of equilibrium and on no other. But what would justify me in believing either of these things?

Can I in fact even *reach* philosophical equilibrium once the above considerations have occurred to me? When they have occurred to me should I not then say something like this to myself: "Why, as far as the warrant or justification belonging to my point of equilibrium is concerned, I might as well have adopted it by opening a book that listed a thousand mutually inconsistent points of philosophical equilibrium and choosing one of them at random. Someone who did that would certainly not have been justified in thinking that the point of philosophical equilibrium he occupied was true. I say that I 'might as well' have done that, but it is no mere fanciful metaphor to say that 'that' is pretty much what nature and nurture and fortune have done *with me*. The point of philosophical equilibrium I occupy depends (perhaps) on predispositions to belief inherent in my genes, (very likely) on what my parents taught me about morals and politics and religion when I was a child, and (certainly) on what university I selected for graduate study in philosophy, who my departmental colleagues have been, the books and essays I have read and haven't read, the conversations I have had at APA divisional meetings as a result of turning right rather than left when I was wandering aimlessly about at a reception. . . . Other philosophers have reached different points of philosophical equilibrium simply because these factors have operated differently in the course of the formation of their opinions. These reflections suggest—and the suggestion is very strong indeed—that I ought to withdraw from the point of philosophical equilibrium I occupy and become a sceptic about the answers to all or almost all philosophical questions."

Well, enough. I have raised certain difficulties for Lewis's views on philosophical method. I have nothing to say about how someone who holds these views should respond to the difficulties I have raised, and nothing to say about how Lewis might have responded to them. As I see matters, very similar difficulties face

13. Members of the Flat Earth Society occupy points of geomorphic equilibrium (so to call them) different from mine. They are aware of all the arguments I could give for the earth's being spherical, and they have ingenious (one has to admit) "refutations" of these arguments. I nevertheless believe that the arguments I can give for the earth's being spherical confer warrant on my belief that the earth is spherical. I answer the challenge to my belief (the one about warrant) that is presented by the existence of other points of geomorphic equilibrium by saying simply that I am better at evaluating arguments concerning the shape of the earth than are the occupants of the other points. I am not willing to make the corresponding response to the challenge to my beliefs presented by the existence of other points of philosophical equilibrium than mine.

anyone who proposes *any* philosophical methodology—provided only that one feature of that methodology is that, whatever other goals philosophy may have, one of its goals is to make true philosophical statements. These difficulties are raised by the fact of pervasive and irresoluble philosophical disagreement. What I have tried to do is simply to suggest that Lewis's philosophical methodology (or his epistemology of philosophy, or whatever it should be called) seems unable to provide its adherents with any reason to suppose that it is epistemically permissible to believe that thoughts are brain processes, that causal relations supervene on the spatio-temporal distribution of local qualities, or that free will is compatible with determinism . . . or to accept any substantive philosophical thesis whatever. Lewis's thesis about the goals of philosophy may be superior to some other theses about the goals of philosophy. It may be superior to any thesis that entails that the primary task of philosophers is to search out knock-down arguments for and against philosophical propositions. But it is no more able than any other such thesis to explain how (in light of the fact of pervasive and irresoluble philosophical disagreement) anyone can be justified in believing anything of philosophical consequence.

3

I will set the difficulties to which Section 2 was devoted aside. I will assume Lewis is right about what philosophers should be aiming at and look at how what he says about free will looks from the point of view provided by his theory of the aims of philosophy.

Lewis believes that free will is compatible both with indeterminism (that is, he believes that a free act can be an undetermined act) and with determinism. It is with the latter belief that I shall be concerned. The most interesting and original aspect of Lewis's compatibilism—presented in his classic essay "Are We Free to Break the Laws?"[14]—is his response to the standard argument for incompatibilism. Versions of this argument (or at least vague intuitions whose articulation would issue in something like this argument) are as old as philosophical concern with free will, but it was not till the 1960s and 70s that the argument was carefully formulated.[15] In those years, David Wiggins, Carl Ginet, Charles Lamb, and I formulated versions of what I am calling the standard argument. I will consider only

14. David Lewis, "Are We Free to Break the Laws?" *Philosophical Papers, Vol. II* (New York and Oxford: Oxford University Press, 1986), pp. 291–98. The paper first appeared in *Theoria* 47 (1981): 113–21. Citations are from *Philosophical Papers II*.

15. There is one important exception to this generalization: C. D. Broad's Knightsbridge inaugural lecture, "Determinism, Indeterminism, and Libertarianism." This lecture (as far as I know) first appeared print in Broad's collection *Ethics and the History of Philosophy* (London: Routledge and Kegan Paul, 1952). It must have been composed in the 1930s, however, since Broad became Knightsbridge professor in 1933. If this lecture had received the attention it deserved, discussion of the problem of free will would have emerged from a very long period of sterile, textbook exchanges thirty years earlier than it did.

my own version of the argument, since that is the one Lewis discusses.[16] The argument, in the form in which Lewis discusses it, turns on the notion of "being able to render a proposition [a proposition that is in fact true] false." The argument begins with the story of a judge (J) who did not raise his hand at a certain moment (T) when his doing so would have prevented a prisoner from being put to death. I claimed to be able to derive the following consequence from the conjunction of this story with determinism: J was not able to raise his hand at T.[17] (That is: "J lacked the ability to raise his hand in such a way that the rising of his hand would have been *complete* at T (and at no earlier moment)"; this statement must not be confused with the following statement: "J lacked the ability to raise his hand in such a way that the rising of his hand would have *begun* at T.") Determinism, I said, implies the following thesis: If P_0 is a proposition that describes the state of the world (in every detail, however minute) at some moment in the remote past and if L is the conjunction of all laws of nature into a single proposition, then the conjunction of P_0 and L entails every truth. I then set out an argument whose conclusion was

If determinism is true, J was not able to raise his hand at T.

This conclusion follows (by sentential logic) from the six premises of the argument. Lewis's discussion of the argument is concerned entirely with two of its six premises:

(5) If J was able to render the conjunction of P_0 and L false, J was able to render L false.
(6) J was not able to render L false.

Lewis contends that one or the other of these premises is false (or would be false in the circumstances imagined); *which one* is false will depend on what is meant

16. That is, the argument as it was presented in "The Incompatiblity of Free Will and Determinism," *Philosophical Studies* 27 (1975): 185–199. I presented a rather different version of the "standard argument" in "A Formal Approach to the Problem of Free Will and Determinism," *Theoria* XL (1974): Part 1, pp. 9–22. (Lewis cites the latter essay, but does not explicitly discuss the version of the standard argument it contains.) Both versions of the argument appeared, with minor revisions, in Chapter IV of *An Essay on Free Will* (Oxford: Clarendon Press, 1983), which also contains a third version of the argument.

17. In my statement of the argument (and in Lewis's discussion of it) the consequence is stated in these words: J could not have raised his hand at T. In both "The Incompatibility of Free Will and Determinism" and *An Essay on Free Will*, I expressed the idea of ability by using "can" and "could have." I now know that the use of "could have" in discussions of the free-will problem is liable to create confusions in the minds of some philosophers (David Lewis not among them), owing to the fact that these words can mean both "was able to" and "might have." (For a discussion of these confusions, see my review of Daniel Dennett's *Elbow Room* in *Noûs* 22 (1988): 609–618.) In more recent writings on free will, I have made it a policy always to use "is able to" instead of "can" in present-tense ascriptions of ability and "was able to" instead of "could have" in past-tense ascriptions of ability. I have decided to adhere to this policy in the present essay, despite the fact that this decision entails changing both the wording of the argument Lewis discusses and the wording of his discussion.

by "is able to render *p* false." (Why does the argument contain this odd form of words, anyway? For the following reason. Determinism is a thesis about the logical relations that hold among certain propositions—those that are laws of nature and those that assert that, at a specified time, the world is in a certain "total state." If one is to investigate the question of the compatibility of determinism and the thesis that human beings are sometimes able to act otherwise than they do, one will need some way to describe an agent's abilities in terms of the agent's power "over" the truth-values of propositions. Infinitival constructions like "is able to raise his hand" do not satisfy this need, nor do most other ordinary idioms. I coined the form of words "is able to render *p* false" to satisfy it.)

We can, Lewis says, distinguish a *weak* and a *strong* sense in which one may be said to be able to render a proposition false. If "was able to render false" (let us call this the Suspect Phrase) is understood in the weak sense, Lewis contends, the compatibilist should deny premise (6) of the argument. (That is to say, on the weak interpretation of the Suspect Phrase, a free agent in a deterministic world is able to render L false.) And if the Suspect Phrase is understood in the strong sense, the compatibilist should deny premise (5). These contentions are perfectly correct, and I will not explicitly discuss the strong and weak definitions of the Suspect Phrase or Lewis's arguments concerning the consequences of these definitions for the truth-values of premises (5) and (6). One might of course want to ask what should be made of the fact that compatibilism entails that there are possible circumstances in which an agent is able to render L false (in *any* plausible sense of "able to render false"). I will, in fact, presently raise just this question, but in relation to my own definition of the Suspect Phrase (which is not the same as either of Lewis's definitions). It is, as Lewis notes, open to me to define the Suspect Phrase in any way I like. He says (rightly), "It does not matter what 'could have rendered false' means in ordinary language; van Inwagen introduced the phrase as a term of art. It does not even matter what meaning van Inwagen gave it. What matters is whether we can give it any meaning that would meet his needs—any meaning that would make all his premises defensible without circularity" (p. 296). (Note that the Suspect Phrase does not appear in the conclusion of the argument.)

In "The Incompatibility of Free Will and Determinism," I did not define the Suspect Phrase. Instead of providing a formal definition of "can render false," I simply gave a few instructive examples (I hoped they were instructive) of cases in which an agent was (or was not) able to render some specified proposition false.[18] In *An Essay on Free Will*, however, I did define "can render false." *An Essay on Free Will* was not yet published when Lewis wrote "Are We Free to Break the Laws?" Apparently, however, I had communicated the definition that was to appear in the book to him in a letter. He discusses this definition in a footnote.[19]

18. Lewis rightly notes the weakness of this method (p. 297, final paragraph): if one attempts to support a principle *p* by giving examples, one's examples may (at best) support only a principle of the form "if *q* then *p*," where *q* is some contingent truth that happens to be true in all the states of affairs laid out in one's examples.

19. This definition was not constructed to block Lewis's argument. It was the outcome of some discussions I had been having with Mark Heller (then a graduate student). Heller had convinced me of the importance of providing an explicit definition of "is able to render *p* false."

This is the definition (as Lewis formulates it—but I have substituted "was able to" for Lewis's "could have"):

> An agent was able to render a proposition false if and only if he was able to arrange things in a certain way, such that his doing so, together with the whole truth about the past, strictly imply the falsity of the proposition. (p. 296, n. 5)[20]

If I understand Lewis, his position is that if I use this definition, the meaning I thereby provide for "was able to render false" is not a meaning that makes all my premises "defensible without circularity." In that case, he maintains, the problematical premise is (6). Lewis says,

> On this definition, Premise 6 simply says that I was not able to arrange things in any way such that I was predetermined not to arrange things in that way. It is uninstructive to learn that the soft determinist [the determinist who believes that people are sometimes able to do otherwise] is committed to denying Premise 6 thus understood. (loc. cit.)

These words are harder to understand on the third or fourth reading than they seem to be on the first. If, as I say, I understand Lewis, he is implying that if the Suspect Phrase is defined in the way I have proposed, then my argument for incompatibilism is circular or begs the question against the compatibilist or something of that sort. (I have never been able to get very clear about what a circular argument is—or begging the question either. I am reminded of a remark that Roderick Chisholm once made in response to a charge of having committed one or the other of these offenses: "I seem to have been accused of the fallacy of affirming the antecedent.") If we are to evaluate this charge, it will be useful to have an unambiguous statement of the argument to which it applies. Let us use the name "the Fully Explicit Argument" for the argument of "The Incompatibility of Free Will and Determinism," modified as follows: all occurrences of the Suspect Phrase are removed from the argument by replacing them with the *definiens* I have proposed. The charge Lewis has made may now be put in these words: the Fully Explicit Argument is circular (or begs the question against the compatibilist). Let us see what we can make of this charge.

In the above quotation, Lewis points out (at least this is very close to what he points out) that the following entailment holds:

> L conjoined with the whole truth about the past strictly implies every truth (that is to say: everything that is so is predetermined to be so)

and

20. Lewis's strict implication is my entailment. Since the quotation from Lewis has introduced the term "strictly implies" into the discussion, and since theses about entailment are going to turn up very frequently in the sequel, I will sometimes myself use "strictly implies," simply to avoid inelegancies like "That this entailment holds entails. . . ."

> J did not raise his hand at T but was able to raise his hand at T

jointly entail

> J was able to arrange things in a way such that he was predetermined not to arrange things in that way.

Let us call this the Trivial Entailment (since it is "uninstructive" to be informed that it holds). Let us give the name 'the Antecedent' to the conjunction of determinism ("L conjoined with the whole truth about the past strictly implies every truth") and the thesis that J did not raise his hand at T but was able to raise his hand at T.

Now consider the following argument for incompatibilism, which takes the Trivial Entailment as its starting-point:

> Here is another way of stating the Trivial Entailment: the Antecedent entails
>
> > J was able to arrange things in a way such that L conjoined with the whole truth about the past strictly implies that he did not arrange things that way.
>
> It follows that the Antecedent entails
>
> > J was able to arrange things in a way such that his arranging things in that way conjoined with the whole truth about the past strictly implies the falsity of L.
>
> And no one has this ability. No one could *possibly* have this ability. But the Antecedent entails that J has this ability, and the Antecedent is a possible state of affairs if free will and determinism are compatible. It follows that free will and determinism are not compatible.

This argument and the Fully Explicit Argument do not, I think, differ in any important way. The two arguments are simply two formulations of the standard argument for incompatibilism. In aid of our attempt to understand Lewis's charge of circularity or question-begging, let us ask what Lewis would have said about this second argument (or this second formulation of the standard argument)—for if the argument in the passage I have quoted from Lewis's footnote shows that the Fully Explicit Argument is circular, it certainly shows that this little argument (to which it is more directly applicable) is circular. Well, he would certainly have rejected one of its premises. He would certainly have rejected *this* premise: "No one could possibly have this ability." Well and good: if he doesn't believe it, he doesn't believe it and the obvious thing for him to do is to say so. What interests me just at present is not whether this premise is true but why I should be accused of circularity or question-begging because I have employed it. (I don't see what *other* premise of the argument such an accusation could be based on.) Is *any* argument for incompatibilism that contains this premise *ipso facto* a "circular" argument? Do I, simply in virtue of employing this premise, "beg the question" against the compatibilist? If so, what would the general lesson be: that no one, in offering a (valid) argument for not-*p*, may include in that argument a premise such that *p*-

ists, on examining the argument, would seize on it and say, "That's the premise that we reject"? That seems to me to be a rather extreme thesis. It seems evident to me that an acceptable philosophical argument may include such a "crucial" premise, and, I think, if there are any good philosophical arguments, most of them do. Most carefully formulated philosophical arguments contain one premise such that (when all terms of art have been unambiguously defined) *that* is the premise philosophers will want to argue about.

It should be evident both from what I said in Section 1 and what I said in the preceding paragraph that I do not regard either the Fully Explicit Argument or the little argument based on the Trivial Entailment or any other argument for incompatibilism as a knock-down argument. Lewis and I agree that the search for knock-down arguments in philosophy is an unrealistic goal. But if that is so, need Lewis pay any attention to any of these arguments? Of what interest could they be to him and to those who share his views on free will? Well, let us look at the present dialectical situation—the dialectical situation in which Lewis and I find ourselves when I have said that it is obviously impossible for one to be able so to arrange things that one's so arranging them together with the whole truth about the past strictly implies the falsity of L, and he has pointed out that compatibilism entails that this *is* possible—in the terms provided by Lewis's own theory of the goals of philosophy. What my argument enables compatibilists like Lewis to do is (in Lewis's words) to measure the price. The cost of compatibilism, or part of the cost, is this: the compatibilist, the philosopher who believes in the possibility of free agents in a deterministic world, must believe that a free agent in a deterministic world is able to arrange things in such a way that one's so arranging them, together with the whole truth about the past, strictly implies the falsity of at least one law of nature. Recall Lewis's words:

> Argle has said what we accomplish in philosophical argument: we measure the price. Perhaps that is something we can settle more or less conclusively. But when all is said and done, and all the tricky arguments and distinctions and counterexamples have been discovered, presumably we will still face the question which prices are worth paying, which theories are on balance credible, which are the unacceptably counterintuitive consequences and which are the acceptably counterintuitive ones. On this question we may still differ. And if all is indeed said and done, there will be no hope of discovering still further arguments to settle our differences.

The question confronting Lewis and me is not, or *should* not be, whether, in employing either the premise "No one could possibly have this ability" or premise (6) of the Fully Explicit Argument—

> J was not able to arrange things in any way such that his doing so, together with the whole truth about the past, strictly implies the falsity of L—

I render my argument circular or beg the question. The question should be: How plausible are these premises? Lewis thinks they are implausible. His reason for

thinking them implausible seems to be this: He finds compatibilism very plausible indeed and he sees that the compatibilist must deny them. He finds this price, these things that the compatibilist must deny, to be, if counterintuitive, at any rate "acceptably" counterintuitive. I find the denials to which the compatibilist is committed implausible, "unacceptably counterintuitive." He and I have, therefore, reached different points of philosophical equilibrium (or we have reached different points of philosophical equilibrium as regards one philosophical question: What should one believe about the standard argument for the incompatibility of free will and determinism?).

Is there anything more to say? Well, I could say something about why I find "No one could possibly have this ability" and premise (6) of the Fully Explicit Argument plausible. I could and I will.

Suppose that Elijah, who is currently in Jerusalem, claims that he is able to be in Babylon ten minutes from now. Suppose further that we, his audience, are able to convince him that the laws of nature and the whole truth about the past together strictly imply that he will not be in Babylon ten minutes from now. Then, surely, Elijah must either withdraw his claim to be able to be in Babylon ten minutes from now or else claim to be able to perform a miracle—for that is what his being in Babylon ten minutes from now would be if the past and the laws of nature together entail that he is not going to be in Babylon ten minutes from now: a miracle.

It is entirely plausible, it is unexceptionable, to define a miracle as an event or state of affairs whose occurrence would be inconsistent with the whole truth about the past and the laws of nature. It would be a mistake to insist that a miracle should be defined as an event whose occurrence would be inconsistent with the laws of nature *tout court*. (Imagine this exchange: "I can perform miracles. I am, for example, able to be in Babylon ten minutes from now." "Oh, that wouldn't be a miracle. A miracle is an event that contradicts the laws of nature. And your being in Babylon ten minutes from now is consistent with the laws of nature, for the laws of nature don't have anything to say about who is where when." It would be a miracle, though.) It is, therefore, entirely plausible to define the *ability* to perform a miracle as the ability to bring about an event or state of affairs whose occurrence would be inconsistent with the whole truth about the past and the laws of nature.

If I had proposed these definitions in an essay that was not about free will but about, say, the concept of the miraculous, no one would have taken exception to them. But these plausible definitions have the following consequence. The ability that premise (6) of the Fully Explicit Argument says that J does not have is the ability to perform a miracle. And since it's entirely plausible to suppose that ordinary people in ordinary circumstances are not able to perform miracles, it's entirely plausible to suppose that (6) is true.

This was my promised argument for the plausibility of (6). It is not, I concede, a knock-down argument for (6) or even for the plausibility of (6). It simply sets forth a price that the compatibilist must pay. Free will in a deterministic world—the argument demonstrates—strictly implies the ability to perform miracles. The compatibilist believes that there are deterministic worlds in which agents have free

will; the compatibilist must therefore grant that in all such worlds, all free agents are able to perform miracles. The compatibilist must grant that, in a deterministic world, freedom is freedom to break the laws. (It is therefore the compatibilist and not the incompatibilist who believes in the possibility of "contra-causal freedom.") But, if my experience of compatibilists is to be trusted, the compatibilist will regard the price as worth paying. Indeed the *soft determinist* (the compatibilist who believes that the actual world is one of the worlds in which determinism and freedom co-exist) will no doubt find the price worth paying. I would expect the typical soft determinist to say something along these lines: "That all free agents are able to perform miracles is perhaps a counterintuitive consequence of soft determinism, but it's an *acceptably* counterintuitive consequence. (And, anyway, it's a consequence only on your definition of 'able to perform a miracle'; no doubt other definitions are possible.) Accepting this consequence is a price worth paying. The price is worth paying because it's just evident that we *are* free and *are* determined (at least for all practical purposes; quantum indeterminacy obviously plays no part in the causal genesis of human action). Note that this acceptably counterintuitive consequence does not entail that anyone is a miracle-worker, that any agent ever does so arrange things that the whole past and the laws of nature together strictly imply that things are not arranged that way. For in every case in which an agent is able so to arrange things, it will be determined that the agent not act on that ability."

My arguments for incompatibilism are therefore like almost all other philosophical arguments: they are not knock-down arguments. They are not arguments that will force the compatibilist to become an incompatibilist on pain of irrationality or cognitive dissonance. They are not arguments that have the enviable property imagined by Robert Nozick: anyone who understands their premises and does not accept their conclusion will *die*. But they are not, on that account, arguments that have no power to affect people's opinions. They are not simply "feel good" arguments for incompatibilists, arguments that incompatibilists can call to mind when they feel their incompatibilist faith flagging. They have in fact demonstrated that they have the power to form philosophical opinion: they have convinced some philosophers who were trying to decide whether to be compatibilists or incompatibilists to become incompatibilists.[21]

21. I think it is very probable that "they have convinced some philosophers" is a gross understatement. I think it is very probable that they have convinced a great many philosophers. Speaking at a conference on free will in the early nineties, I made a remark to the effect that compatibilism was the standard view among philosophers. Michael Slote, who was in the audience, said that he thought that, on the contrary, incompatibilism had become the standard view, or at least the majority view. A few years later, I asked Ted Warfield whether he thought that was right. Warfield, who comes as close as is humanly possible to knowing what every analytical philosopher thinks about anything and is very knowledgeable indeed about the ins and outs of the free-will controversy, replied that he thought that the majority of analytical philosophers who had actually worked on the free-will problem were incompatibilists, and that the majority of analytical philosophers (full stop) were compatibilists. If it is indeed true that the majority of analytical philosophers who have actually worked on the free-will problem are incompatibilists, a very large part of the explanation of this fact lies in the influence of the various versions of the "standard" argument for the incompatiblity of free will and determinism on philosophers who were graduate students in the seventies and eighties.