Some Thoughts on An Essay on Free Will

By Peter van Inwagen

It has been just over thirty years since the publication of *An Essay on Free Will*.[1] In this essay, I record some thoughts I have had at various points during those thirty years about the book, its reception, and the way analytical philosophers have thought about the free-will problem since its publication.[2]

I will not summarize the book. Nor will I be concerned to defend its arguments—or at least not in any very systematic way. I will instead present some thoughts on three topics:

1. The question 'If I were to revise the book today, if I were to produce a second edition, what changes would I make?'
2. Aspects of the book I should like to call to the attention of readers (aspects that, in my view, readers of *An Essay on Free Will*, have been insufficiently attentive to).
3. The course of the discussion of the problem of free will subsequent to the publication of the book.

If I were to revise the book today, what changes would I make?

First, I would use an entirely different vocabulary to frame the problem to which the book is addressed. In the paragraphs that follow I will describe this "different vocabulary," and I shall try to explain the reasons for my present dissatisfaction with certain of the words and phrases that figured prominently in *An Essay on Free Will*.

The most salient change I would make, although perhaps not the philosophically most important one, is that I would not now use the phrase 'free will'. In fact, I would not use even the adjective 'free'—I would not speak of free actions, free agents, or free choices. Nor would I use the adverb 'freely' and the noun 'freedom'. In my view, these words have little meaning beyond that which the philosopher who uses them explicitly gives them, and yet philosophers persist in arguing about what they do or should mean. They enter into disputes about

Peter van Inwagen is the John Cardinal O'Hara Professor of Philosophy in the University of Notre Dame Department of Philosophy. His areas of research interest are metaphysics, philosophy of theology, and philosophy of action. He has delivered a number of important named lectures, was elected to the American Academy of Arts and Sciences in 2005, served as President of the Central Division of the American Philosophical Association in 2008-2009, and currently is President of the Society of Christian Philosophers. He has published a number of books, focusing mostly on issues in philosophy of religion.

what "free will" and "free choices" and "acting freely" and "freedom" *really are*. These philosophers have fallen prey to what I may call *verbal essentialism*. That is to say, it is essential to their discussions that they involve certain *words*: 'free', 'freely', 'freedom'. … It would be impossible to translate their discussions into language that did not involve those words. The essential content of *An Essay on Free Will*, however, could have been presented without using 'free' or 'freely' or 'freedom'. (I have shown in detail how to do this in an as-yet-unpublished paper called "The Problem of Fr** W*ll.") In this essay, however, I will use the phrase 'free will' (and 'free' and 'freely') simply because there is no readily available alternative. This phrase should be understood in the sense I gave it in *An Essay on Free Will*. (Or see the definition of 'free act' in footnote 10 below.)

I would, moreover, not use the phrase 'could have'—and I would be particularly careful to avoid the phrase 'could have done otherwise'. 'Could have' is grammatically ambiguous, and this has caused a great deal of confusion in discussions of the free-will problem in English. (As far as I know, there is no parallel ambiguity in any other language.[3]) Sentences of the form 'X could have done Y' can mean either 'X might have done Y' (i.e., 'This is how things might have turned out: that X did Y') or 'X was able to do Y'. The ambiguity is nicely brought out by the following example, which I adapt from Austin. A corrupt public official says to a subordinate, 'You could have exposed me this morning'. Here are two ways in which she might continue the sentence: (i) '…for God's sake, be careful about what you say when you're talking to the press'; (ii) '…and you didn't. I want you to know that I'm grateful'. If she had said (i), 'You could have exposed me' would have meant 'You might have exposed me'; if she had said (ii), 'You might have exposed me' would have meant 'You were able to expose me' or 'You were in a position to expose me' or 'It was within your power to expose me'. In almost all cases, when the phrase 'could have done otherwise' is used in discussions of the free-will problem, its intended meaning is 'was/were able to do otherwise'. But discussions of free will generally involve mention of determinism and indeterminism, and to say that one's action was undetermined is to say that one "could have done otherwise" in the *other* sense of the phrase. One of the confusions that has resulted from the double meaning of 'could have done otherwise' is that some critics of libertarianism[4] have supposed that when libertarians say (for example), 'She was not morally responsible for what she did because she could not have done otherwise' they mean 'She was not morally responsible for what she did because her act was determined to occur'. Now libertarians do believe that 'X was able to do Y' entails 'X might not have done Y' (i.e, 'The world as it was just before X did Y *might have* evolved in such a way that X did not do Y'), but they regard this as a substantive philosophical thesis. They do not regard 'X was able to do Y' and 'X might not have done Y' as two ways of saying the same thing.

For an extended example of the confusions generated by the ambiguity of 'could have' see Chapter 6, "Could Have Done Otherwise," of Daniel Dennett's *Elbow Room*.[5] In a revised version of the book, I would replace almost all occurrences of 'could have done' with 'was/were able to do'.

In connection with this replacement, I would say more than I did (and I did say a lot) about the sense of 'able to' that is relevant to the free-will problem—

for the phrase has more than one sense (it may not be grammatically ambiguous like 'could have' but it *is* ambiguous). Suppose, for example, that Martha Argerich is stranded on a desert island (where, of course, there is no piano). Is she able to play *Pictures at an Exhibition*? In one sense of 'able to play' she is (she knows it, as the idiom has it, forwards and backwards), and in another sense she is not. I would explain the relevant sense of 'able' in terms of what is presupposed by making a promise: if a fellow castaway begs Argerich to play *Pictures* (then and there), she is not in a position to promise to do so, for (in the sense of 'able' that is relevant to the problem of free will), she would not be *able* to keep that promise.

In the revised book, I would not use the phrase 'moral responsibility'—for, in my view, this phrase is used in current philosophy without any clear sense. I would replace all references to 'moral responsibility' with references to fault or blame. (To *moral* fault or blame.) And I would speak of fault and blame only in connection with the consequences of actions (or failures to act). Thus, I would not say anything like 'Alice is morally responsible for telling lies to Frank'—for it is not at all clear what that means. I would instead say things like, 'Frank's unhappiness is Alice's fault' or 'Alice is to blame for Frank's unhappiness'. (Alice's lies would come in at a later stage in a conversation about Alice and Frank: "*Why* is Alice to blame for Frank's unhappiness?" "Because his unhappiness is due to the scurrilous lies she told him about his wife.")

In the revised book, I would replace "Principle β" with a different principle. Principle β was:

(1) *p* and no one has, or ever had, any choice about that
(2) If *p*, then *q*, and no one has, or ever had, any choice about that *hence*,
(3) *q* and no one has, or ever had, any choice about that.

Or, in abbreviated form,

Np, $N(p \rightarrow q) \rightarrow \rightarrow Nq$.

And Thomas McKay and David Johnson have shown that β is invalid.[6] Consider a coin that is never tossed. Suppose that I have a choice about whether the coin is tossed, but that neither I nor anyone else has, or ever had, a choice about how the coin would fall if tossed. Let '*Notheads*' abbreviate 'The coin never falls heads' and '*Nottails*' abbreviate 'The coin never falls tails'. It is obvious that

(1) N(*Notheads* → (*Nottails* → (*Notheads* & *Nottails*))),

since '*Notheads* → (*Nottails* → (*Notheads* & *Nottails*))' is a logical truth. The following two statements are also obvious:

(2) N*Notheads*
(3) N*Nottails*,

owing to the fact that *Notheads* and *Nottails* are both true and no one has a choice about how the coins would fall if tossed.

If β is valid, then (1) and (2) imply

(4) N(*Nottails* → (*Notheads* & *Nottails*)),

and (3) and (4) imply

(5) N(*Notheads* & *Nottails*).

And (5) is false, since I have a choice about whether the coin is tossed: if I choose not to toss the coin (my actual choice), *Notheads* & *Nottails* will be true. If I choose to toss the coin, *Notheads* & *Nottails* will be false, since the coin must fall either heads or tails.

The reason I mistakenly supposed that β was valid was this. I mistakenly supposed that the only way in which it could be that one had no choice about the truth-value of a proposition would be for the truth-value of that proposition to be in some way so firmly "fixed" that one was unable to change it. I did not see that that there is another way for one to have no choice about the truth-value of a proposition: for that truth-value to be a mere matter of chance.

I would now formulate β differently. Or, if you like, I would substitute another principle for β, a principle that does not contain the phrase 'has no choice about'. I propose the following. Say that it is *a humanly unalterable truth* that *p* just in the case that *p* and nothing that any human being is or ever has been able to do is such that if someone were to do it, that person's action might result (could possibly result) in its not being the case that *p*. "Revised β" would then be

(1) It is a humanly unalterable truth that *p*
(2) It is a humanly unalterable truth that if *p*, then *q*
hence,
(3) It is a humanly unalterable truth that *q*.

I believe this principle to be valid and I believe that the premises of the version of the Consequence Argument that employed β would all be true if each occurrence of 'N' in that argument were replaced with 'it is a humanly unalterable truth that'.

If I were to revise *An Essay on Free Will*, I would change what I said about psychological laws. What I said was this (pp. 63–4):

> Suppose psychologists discover that no one who has received moral training of type A in early childhood ever spreads lying rumours about his professional colleagues. Suppose you and I in fact received such training. Does it follow that we *can't* engage in this odious activity? I don't see why it should be supposed to follow. … Suppose further that you and I are in fact *able* to spread lying rumours about our colleagues. Does it follow that a statement of the regularity we have supposed psychologists to have discovered is, though true, not a law? [I do not see why it should not be regarded as a law.] "But why", someone may ask, "does this regular pattern of behavior occur if people don't *have* to conform to

> it?" Note that the only people in a position to depart from it are those who have in fact had training of type A. Perhaps it is just these people who *see the point* in not spreading lying rumours. To come to see the point in not exercising an ability one has is not to *lose* that ability.

Conversations with Alexander Rosenberg, and some thoughts that I had after reading a science-fiction novel called *Protector* by Larry Niven, gradually convinced me that this passage was radically confused. (The "Protectors" of Niven's novel are beings, all of whom are sterile males formerly capable of reproduction, and each of whom, of biological necessity, ascribes intrinsic value to only one thing: the preservation of his own bloodline. Protectors, moreover, are far more intelligent than any human being, and, in consequence, each Protector almost always sees immediately, in any situation in which he finds himself, what course of action is most likely to preserve his bloodline—"and straightaway he acts." Niven has one of his Protectors say at one point, "Protectors have precious little free will." When I first read those words, my immediate reaction was to smile and to regard them as a rather typical non-philosopher's confusion about free will. A decade or so later, I came to the conclusion that it was I who had been confused.[7]) My reasons for regarding the passage quoted above about psychological laws as confused eventually found expression in a paper called, "When Is the Will Free?"[8] In that paper, I defended the position that the principles—Principle β, for example—that I used to argue for the incompatibility of free will and determinism also support the proposition that if human beings are ever able to act otherwise than they in fact do, this can be the case only very rarely. (The distinction between "Original β" and "Revised β" is not relevant to my defense of this position.) A revised version of *An Essay on Free Will* would incorporate this position.

If I were to rewrite the book, it would contain an extensive discussion of the implications of recent developments in neuroscience for the problem of free will.[9]

Since the publication of *An Essay on Free Will*, it has become increasingly clear to me that free will is a philosophical mystery—something that philosophers do not understand at all. (It is not the only one. For example, no philosopher understands conscious experience or the apparent "passage" of time.) I do not mean to imply that free will is a mystery in the theological sense: something that is beyond all possibility of human comprehension. That may or may not be the case. I contend only that as of this date, no philosopher has achieved an understanding of free will. That may be because free will is indeed something that human beings are incapable of understanding, but it may be because we human beings have not yet discovered the right way to think about free will. I will lay out the essence of this mystery in four fairly simple statements—labeled 'first', 'secondly', 'thirdly' and 'fourthly'. First, there are excellent arguments for each of the following three propositions:

(1) If antecedent conditions and the laws of nature determine the way in which a human being shall act at a certain time, then that person's act at that time is not free.[10] (This proposition, of course, is the proposition commonly called 'incompatibilism'.)

(2) If antecedent conditions and the laws of nature do *not* determine the way in which a human being shall act at a certain time, then that person's act at that time is not free.
(3) If a human being's acts are never free, then the consequences of those acts are not the fault of that human being.

Secondly, the following fourth proposition seems to be true beyond all possibility of dispute.

(4) Some of the consequences of some of the acts of some human beings are their fault.

Thirdly, these four propositions form a logically inconsistent set, and, therefore, either the excellent arguments for at least one of first three propositions must contain some flaw or else it must be that (to take one example among many millions of compelling examples) the deaths of six million Jews in the extermination camps were not anyone's fault. Fourthly, no one knows of even a plausible candidate for a flaw in any of the arguments for the first three of the four propositions (not, at any rate, in the arguments *I'm* thinking of), and to deny the fourth would be simply bizarre. I doubt whether many philosophers will agree with my statement that "no one knows of even a plausible candidate for a flaw in any of the arguments for the first three of the four propositions," but it represents my considered judgment. My own view is that there is a flaw in the argument (the argument that *I* think is the best argument for this conclusion) for the proposition 'If antecedent conditions and the laws of nature do *not* determine the way in which a human being shall act at a certain time, then that person's act at that time is not free'. But I haven't any idea what this flaw might be.

I would recommend that the "problem of free will" be understood as follows: it is the problem of discovering a flaw in at least one of the arguments for the first three propositions—or else of explaining how the seemingly self-evident fourth proposition, could, despite all appearances, be false. In my judgment, no one has the least idea how to solve this problem. That is what I mean by saying that "free will is a mystery." If I were to revise *An Essay on Free Will*, I would give the thesis that free will is a mystery a very prominent place in the revised work (although, as I have said, in stating and defending this thesis I would not use the words 'free will').

In my view, few if any of my fellow "libertarians"—that is, incompatibilists who accept the reality of free will—appreciate the immense power of the *Mind* Argument (the conclusion of which is the second of the propositions above: the proposition that undetermined human actions cannot be free). One typical reaction of libertarians to that argument is to declare that the problem it poses for libertarianism can be solved by positing that some events are caused not by earlier events but by *substances*, to wit, human agents. These libertarians hold that the *agent* (as opposed to some event that occurs within the agent) is sometimes the cause of the agent's actions. I have since presented arguments for the conclusion that even if "agent causation" indeed exists, it is irrelevant to the problem that

the *Mind* Argument poses for libertarianism.[11] I would include these arguments in a revised version of the book.

Much of what was said about quantum mechanics in section 6.2 of the book was out of date even when the book was written—although, of course, I did not realize that at the time; I did not realize that several of the most important of my statements about quantum mechanics were based on obsolete sources (this is particularly true of my statements about von Neumann's "proof" of the impossibility of supplementing quantum mechanics with "hidden variables"). I would completely re-write section 6.2.

Many philosophers writing on free will suppose that "libertarian free will" and "compatibilist free will" are two different things. (This is particularly true of compatibilists.) They suppose, that is, that libertarians believe in a kind of free will that is incompatible with determinism, and that compatibilists believe in a kind of free will that is compatible with determinism. They suppose that libertarians reject compatibilism on the ground that to give the name 'free will' to what the compatibilists call by that name is to offer (in Kant's words) "a miserable substitute"—that in so using 'free will' they lead the unwary reader into (in James's words) "a quagmire of evasion."[12] And they suppose that compatibilists reject libertarianism on the ground that what libertarians give the name 'free will' to does not exist (and is, moreover, something that a sane person should neither believe in nor want to have). In my view, this position is simply false and allegiance to it has been the occasion of an immense amount of unclarity and confusion in the literature on free will. (Of course this confusion would simply vanish if, as I recommend, philosophers were to cease to use such phrases as 'free will' and 'free agent' and 'free act'.) If I were to revise *An Essay on Free Will*, I would include a presentation of my reasons for holding this view.[13]

Aspects of the book that readers have been insufficiently attentive to

One frequently hears variants on the following challenge to libertarianism—as frequently today as thirty years ago:

> If free will were, as libertarians contend, incompatible with determinism, one could never know whether one—whether *anyone*—had free will unless one knew whether the laws that governed the world were deterministic. And if (as libertarians believe) ascriptions of fault or blame presuppose the existence of free will, then no one could know whether anything was anyone's fault unless one first knew that determinism was false. And they have no good reason to believe that determinism is false. (Quantum mechanics provides no such reason. The indeterministic "collapse" of the wave function is a feature only of one *interpretation* of quantum mechanics. And, in any case, it is doubtful whether events at the quantum level play any role in human action.) Libertarians should therefore be skeptics about whether anything has ever been anyone's fault. But libertarians seem to be perfectly confident that certain states of affairs are the fault of certain people, that certain people are to blame for certain things. This is not a consistent position.

I would call the attention of anyone who accepts any argument that is even remotely similar to this argument to section 6.3 of *An Essay on Free Will*. I do not claim to be familiar with the whole of the vast literature on free will, determinism, and moral blame, but to the best of my knowledge, the arguments of that section have not only never been adequately answered, but they have never been discussed at all. In my view, anyone who accepts anything like the above argument should address the arguments of section 6.3.

I turn now to the topic of "Frankfurt counterexamples." One of the propositions that figures in the statement of "the problem of free will" above is:

(6) If a human being's acts are never free, then the consequences of those acts are not the fault of that human being.

Many philosophers suppose that this proposition was shown to be false by Harry Frankfurt in his classic paper "The Principle of Alternate Possibilities."[14] In that paper, Frankfurt presented a counterexample to the "Principle of Alternate Possibilities":

> PAP A person is morally responsible for what he has done only if he could have done otherwise.

If one re-states (6) in terms of "moral responsibility" (and 'person'), one obtains something like

(7) If a person's acts are never free, then that person is not morally responsible for the consequences of any of those acts,

which is equivalent to

(8) A person is morally responsible for some of the consequences of some of his acts only if that person's acts are sometimes free.

If a person's act is free just in the case that that person "could have done otherwise" (that is, was able to do otherwise, was able to do something else or nothing at all; *cf*. *n* 10), then (8) is, more or less, equivalent to

(9) A person is morally responsible for some of the consequences of some of his acts only if that person is sometimes (i.e., at certain points in his life) able to do otherwise.

(The reader will see that I have been attempting to transform (6) into something as similar to PAP as possible.)

It is the burden of the long discussion of Frankfurt's arguments in *An Essay on Free Will* (sections 5.3–5.7) that, even on the assumption that Frankfurt has presented a successful counterexample to PAP, (a) this counterexample is not a counterexample to (9), and (b) it is not possible to use the "general idea" behind Frankfurt's counterexample to PAP to construct a counterexample to (9).

I therefore maintain that if Frankfurt has indeed refuted PAP, this refutation is irrelevant to the problem of free will, or at least irrelevant to the problem I call 'the problem of free will'. But a significant proportion of the writers on the problem of free will continue to treat "Frankfurt counterexamples" as an important contribution to our understanding of the problem of free will. I would direct their attention to sections 5.3–5.7 of *An Essay on Free Will*.[15]

I turn, finally, to the challenge, "But what would you say if determinism, or something that might be called 'determinism for all practical purposes', turned out to be true?" I am not infrequently asked questions whose essential point may be summarized as follows.

> Whatever may be the case as regards the general metaphysical thesis of determinism, it is certainly a very real possibility that human beings are "essentially deterministic systems"—in the sense in which your computer is an essentially deterministic system. (Even if there are all sorts of undetermined events going on at the quantum level inside your computer, it is vastly unlikely that these events will have the consequence that the behavior of your computer will exhibit any indeterminacy at level of "observables"—vastly unlikely, for example, that it is indeterminate what will appear on the monitor, given a precise description of what you have done at the keyboard.) And if determinism is incompatible with the proposition that human beings have free will, then—surely?—the thesis that human beings are essentially deterministic systems is also incompatible with that proposition. What would you say if that possibility were shown by scientific investigation to be realized? What would you say if science produced a convincing demonstration that human beings were essentially deterministic systems?

I have answered this question with some care in section 6.4 of *An Essay on Free Will*, although this answer has received little or no attention from writers on free will. My answer takes the form of an argument for the thesis that the most reasonable response for me to make to such a "convincing demonstration" would be to conclude that Principle β was invalid—or, since β is invalid in any case, that "Revised β" was invalid. "Revised β" seems to me to be an obvious truth—it seems to possess a certain "luminous evidence" (Locke) or to "force itself upon the mind as true" (Gödel). But the history of thought provides a fund of examples of propositions that seemed to *very* able thinkers to have those features and which eventually proved to be false. One may cite Zeno's conviction that every object is motionless at an instant (a landscape painting displays the spatial relations among certain objects at an instant; if you inspect such a painting, you will observe that everything in it is *not moving*), Galileo's conviction that there are more integers than there are odd integers, Frege's appeal to what we now call "the unrestricted comprehension principle" in set theory, and the status of the Galilean law of the addition of velocities in pre-Einsteinian physics. (And, if you are willing to regard *me* as a "*very* able thinker," you may add my former certainty that β—unrevised β—was true to this list of examples.)

The course of the discussion of the problem of free will subsequent to the publication of **An Essay on Free Will**

I believe that the most important contribution to the literature on free will subsequent to the publication of *An Essay on Free Will* was David Lewis's essay "Are We Free to Break the Laws?"[16] (Although this essay appeared in a number of *Theoria* dated 1981, that number of the journal may well not actually have appeared till after the publication of the book. In any case, it was written before Lewis had read the book. The essay is, however, a profound critique of the argument of my own essay "The Incompatibility of Free Will and Determinism,"[17] and therefore a profound critique of one of the three arguments for incompatibilism that are presented in Chapter III of *An Essay on Free Will*.) In my own deeply prejudiced view, "Are We Free to Break the Laws?" is the only publication of real philosophical significance concerning the arguments of the book. (I have replied to it in an essay called "Freedom to Break the Laws."[18])

In my opinion—deeply prejudiced, to be sure—the philosophical literature on free will, subsequent to the publication of my book and Lewis's essay, has degenerated into a sterile scholasticism (in the pejorative sense of the word; I do not mean to imply that scholasticism in the pejorative sense was any more common in the medieval schools than it has been in any other philosophical community). That is, there have been no new arguments or ideas of any real consequence. The parties to the discussion of the problem of free will since 1983 know all the relevant arguments and concepts that pertain to every aspect of the problem, and dispute about those arguments and concepts without saying anything that is both new and important about them. They employ, moreover, a standard or "set" vocabulary for discussing the problem that is in many respects unsatisfactory and of which they are insufficiently critical—or, to be frank, not critical at all. (And this is all the more damaging because the meanings of many of the items in the "set vocabulary" have changed over time, and not for the better—a point to which I will return very shortly.)

I would add two qualifications to this generalization. First, as I said earlier, recent work in neurobiology has seemed to many to have important implications for the question whether human beings have free will, and, in consequence, some neurobiologists have had a great deal to say about this question. Whether any of what they have said is of any great philosophical importance is a difficult question to answer. The jury is still out, as they say. But, philosophically important or not, it is *new*, and the charge of "sterile scholasticism" that I have brought against the work of philosophers writing on free will does not apply to it. (I do not mean to imply that philosophers have been inattentive to the implications of neurobiology for the problem of free will. There has been some very interesting work by philosophers on those implications, and that work, too, must be exempted from the charge of sterile scholasticism.)

Secondly, there have been some developments in the work of philosophers on free will in the last thirty years that I would condemn on grounds other than "sterile scholasticism." I have observed, in examining this work, a phenomenon that I can only describe as "terminological degeneration." One case of this has already been mentioned: the introduction of the phrases "compatibilist free will"

and "libertarian free will" into discussions of the free-will problem. I will mention two other cases—cases not unrelated to each other.

(i) There has been a tendency to use 'free will' in a sense something like this:

> To possess free will is to have open to one alternative possibilities in whatever sense of 'alternative possibilities' it is that is relevant to ascriptions of moral responsibility.

Now I have deprecated both the phrase 'free will' and the phrase 'moral responsibility'. But my reasons for disliking these two phrases are largely irrelevant to my present point. As to 'moral responsibility', if I were to change the definition in (i) to

> To possess free will is to have open to one alternative possibilities in whatever sense of 'alternative possibilities' it is that is relevant to ascriptions of moral fault or blame.

my objections to the "new" definition of 'free will' would be essentially the same. Let us therefore turn to those objections.

In discussions of the free-will problem for at least several decades before (and for at least a few years after) the publication of *An Essay on Free Will*, the phrase 'free will' had an agreed-upon and reasonably precise meaning:

> An agent has free will if he or she sometimes acts freely; and an agent acted freely on a certain occasion if that agent was able to have done something other than what he or she did.[19]

My first objection can be presented as a rhetorical question: Why change the long-agreed-upon meaning a philosophically significant term has had?—the meaning, indeed, it had throughout a significant episode in the history of the problem of free will, to wit the period (1965–1985) during which incompatibilism became a respectable philosophical position?

And here is a second rhetorical question: Surely the new definition is a lot less clear than the "classical" one?

It seems to me, in fact, that the very idea of adopting a new or revised definition of 'free will' rests on a confusion—a failure to recognize the fact that 'free will' is a philosopher's term of art, a purely technical term that does occur in everyday discourse.[20] The attempt to provide a new definition of 'free will' is an example of the above-mentioned "verbal essentialism" that infects much of the current discussion of the relations between determinism, indeterminism, and the ascription of moral blame. It rests on the false belief that the *phrase* 'free will' and the *words* 'free' and 'freely' and 'freedom' are—owing the role that they supposedly play in our everyday discourse about fault and blame—of philosophical significance, the false belief that the use of these particular verbal items is *essential* to the posing and discussion of "the problem of free will." It is this apparently widespread false belief that has led me to the conclusion that it

would be a good thing if 'free will' and 'free' and 'freely' and 'freedom' were entirely eliminated from discussions of philosophical problems concerning fault and blame and the causal antecedents of human action.

(ii) There has been a tendency to change the meaning of 'compatibilism'. The original meaning of this word (I believe that the word was coined and given this meaning by Keith Lehrer in the 1960s) and the meaning it had during the "classical" period of the discussion of free will by analytical philosophers (1965–1985) was

> The existence of free will (the existence of human beings who sometimes act freely in the sense given above) is compatible with determinism (the thesis that at any given moment, the laws of nature permit only one possible future)'.[21]

And this is the "new" meaning of 'compatibilism':

> The existence of *moral responsibility* is compatible with determinism.[22]

(Or, as I should prefer to say, the existence of human agents who can properly be blamed for some of the consequences of their acts is compatible with determinism. But my reasons for disliking the phrase 'moral responsibility' are irrelevant to my present point.) This seems to me to be a harmful terminological innovation for two reasons. First, it has the consequence that the word 'compatibilism' will have been used at two different senses at various places in the free-will literature—and it may not always be clear which of the two senses a given author means it to have. Secondly, it leaves those who use it without a name for the thesis that the existence of free will is compatible with determinism. (Unless, of course, those who insist on using 'compatibilism' in the new sense invent a new word to express the old sense. But surely it would be a better procedure to let 'compatibilism' continue to mean what it originally meant, and to coin some new term—semi-compatibilism or MR-compatiblism or some such—as a name for the thesis that moral responsibility and determinism are compatible? A nice compromise, and one I should have no objection to, would be to drop 'compatibilism' *simpliciter* and to introduce two new terms to express the two senses unequivocally—perhaps 'FW-compatibilism' and 'MR-compatibilism'.)

Now giving the word 'compatibilism' this new sense would be defensible if it were clear that the question, 'Is the existence of moral responsibility compatible with determinism?' could be investigated without raising and attempting to answer the question 'Is the existence of free will compatible with determinism?' And indeed many philosophers think that it is possible to discuss the former question without discussing the latter—possible, that is, if 'free will' in the latter question is understood in its "classical" sense:

> An agent has free will if he or she sometimes acts freely; and an agent acted freely on a certain occasion if that agent was able to have done something other than what he or she did.

(Of course, if 'free will' is understood in *this* sense:

> To possess free will is to have open to one alternative possibilities in whatever sense of 'alternative possibilities' it is that is relevant to ascriptions of moral responsibility,

if the concept of "free will" is defined in terms of its relation to moral responsibility, then by definition any attempt to answer the question, 'Is the existence of moral responsibility compatible with determinism?' will involve an attempt to answer the question 'Is the existence of free will compatible with determinism?')

Many philosophers, then, think that is it possible to investigate the question whether the existence of moral responsibility is compatible with determinism without raising and attempting to answer the question whether the existence of agents who, on some occasions, are able to act otherwise than they have in fact acted is compatible with determinism. That thesis may be true or it may be false—*I* certainly think it's false—but it is a *substantive* philosophical thesis. The appropriation of 'compatibilism' to mean 'the thesis that the existence of moral responsibility is compatible with determinism' seems to suggest that this substantive philosophical thesis has somehow been shown to be true—which it certainly has not been. I therefore contend that the practice of using 'compatibilism' in this new sense be strongly resisted. I insist that the "standard vocabulary" employed in discussions of the problem of free will *must* include a term for the thesis that the existence of agents who, on some occasions, are able to act otherwise than they have in fact acted is compatible with determinism. I recommend (although I do not insist) that that term be 'compatibilism'. I do, however, insist that 'compatibilism' *not* be used as a name for the thesis that moral responsibility is compatible with determinism. If 'compatibilism' is not to be used in its original sense, then it should not be used at all, and, as I have suggested, some new pair of unequivocal terms (such as 'FW-compatibilism' and 'MR-compatibilism') should be devised and adopted.

Notes

[1] Oxford: at the Clarendon Press, 1983.

[2] This essay is adapted from a new preface I have written for the forthcoming French translation of the book by Cyrille Michon.

[3] The ambiguity is rooted in the fact that when the English modal auxiliary 'can' has a personal subject, it does not govern an infinitival clause. Consider, for example, the French verb '*pouvior*'. One way of saying '*Je peux jouer au tennis*' in English is, 'I can play tennis'—and not '*I can to play tennis'. By contrast, the English verb-phrase 'to be able', like '*pouvior*' governs an infinitival clause: 'I am able to play tennis'. The German verb '*können*' and the Latin verb '*posse*' also govern infinitival clauses.

[4] Libertarianism is the conjunction of two theses: (a) that human beings have free will, and (b) incompatibilism, or the thesis that one's having free will is incompatible with one's acts being determined.

[5] *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge: The MIT Press, 1984). I have laid out the consequences of Dennett's confusion of the two senses of 'could have' in "Dennett

on 'Could Have Done Otherwise'," *The Journal of Philosophy* 81 (1984), pp. 565–567. See also my review of *Elbow Room* in *Noûs* 22 (1988), pp. 609–618.

[6] Thomas McKay and David Johnson, "A Reconsideration of an Argument against Compatibilism," *Philosophical Topics* 24 (1996), pp. 113–122.

[7] When I told this autobiographical anecdote to David Lewis, who was familiar with Niven's novel, he understood my point immediately—although he did not agree with it.

[8] *Philosophical Perspectives, Vol 3: Philosophy of Mind and Action theory* (1989), *pp*. 394–422. Available on line at <http://andrewmbailey.com/pvi/>.

[9] Walter Sinnott-Armstrong (ed.) *Moral Psychology, Volume 4: Free Will and Moral Responsibility* (Cambridge, Mass.: MIT Press, 2014) is an excellent introduction to this topic.

[10] Let us say that a person's act was *free* just in the case that (i) that person consciously decided to perform that act rather than some contemplated alternative act(s); and (ii) that person was able to perform at least one of those alternative acts.

[11] "Free Will Remains a Mystery," *Philosophical Perspectives, Vol 14: Action and Freedom* (2000), *pp*. 1–19. Available on line at <http://andrewmbailey.com/pvi/>.

[12] The more usual phrase attributed to Kant in discussions of free will in English is 'a wretched subterfuge'. This was Abbot's translation of '*ein elender Behelf*'. But '*Behelf*' is better translated as 'substitute'. The passage in which Kant uses the phrase occurs in the second chapter of the Analytic of Pure Practical Reason in the *Critique of Practical Reason*. James's almost equally famous 'quagmire of evasion' occurs in "The Dilemma of Determinism."

[13] I would direct any philosopher who thinks that "compatibilist free will" and "libertarian free will" are two different things to David Lewis's marvelous essay "Are We Free to Break the Laws?" (*Theoria* 47 (1981), pp. 113–121. Available on line at <http://andrewmbailey.com/dkl/>. Lewis, with his usual clarity of mind, realized that he, a compatibilist, and I, a libertarian, were talking about the *same thing* when we spoke of agents "acting freely," and that the issue between us was whether this one thing was compatible with determinism. A revised version of *An Essay on Free Will* would certainly include an extensive discussion of "Are We Free to Break the Laws?"

[14] Harry G. Frankfurt, "The Principle of Alternate Possibilities," *The Journal of Philosophy* LXVI (1969), pp. 829–839. (The correct word, by the way, would have been 'alternative', not 'alternate'.)

[15] Those sections are a condensation of a more detailed discussion of "Frankfurt-style examples" in "Ability and Responsibility," *The Philosophical Review* LXXXVII (1978) pp. 201–224.

[16] See above, p. xx.

[17] *Philosophical Studies* 27 (1975) pp. 185–199. Available on line at <http://andrewmbailey.com/pvi/>.

[18] *Midwest Studies in Philosophy* XXVIII (2004), pp. 334–350. Available on line at <http://andrewmbailey.com/pvi/>.

[19] The reader will note that there is a certain tension between this definition and my definition of 'free act' in note 10. That definition reflects my own way of using 'free act'. The definition to which this note is appended should be thought of as an attempt to state the common element of the many very-similar-but-not-quite-identical definitions of 'free will' used by various analytical philosophers in the years 1965–1985.

[20] This bald statement requires two qualifications. (a) 'Free will' occurs in everyday discourse as a component of the longer phrase 'of * own free will', where the asterisk represents the position of a possessive pronoun. This phrase implies nothing but the absence of coercion. To say, for example, that Kim Philby acted as a Soviet agent "of his own free will" means nothing more than his acting as a Soviet agent was uncoerced: he did not so act because he had been threatened with unpleasant consequences if he refused to be a Soviet agent. And, of course, everyone will

agree the question whether a given action was coerced or uncoerced has nothing to do with the questions like 'Is the state of the world at a given time is determined by its states at earlier times?' and 'Does God foresee everything we do?' and 'Do the Libet experiments show that our actions are the result of events prior to our conscious choice to perform them?' (b) Technical terms from various sciences and disciplines ('entropy', 'catalyst', 'evolution'…) do find their way into everyday discourse, where they are used without any real understanding of their technical senses—impressionistically, as it were. And, of course, this has happened with 'free will'. If a popular science writer tells her readers that the Libet experiments prove that our belief in "free will" is an illusion, most of her readers will think that they have know what it was that she has said has been proved to be an illusion—despite the fact that if you asked them to say *what* our belief in free will was a belief in, they could give no answer that had any intelligible content.

[21] Two propositions are "compatible" if it is metaphysically possible for them both to be true. Thus 'compatibilism' was originally a name for the thesis that there are possible worlds in which (a) at any given moment, the laws of nature permit only one possible future and (b) there are agents who sometimes act freely. Of course, there would be little point in one's being in that sense a compatibilist if one did not accept the following conditional thesis: if the *actual* world is "deterministic," then it is one of the deterministic worlds in which there are agents who sometimes act freely; but that thesis is not strictly speaking implied by "classical compatibilism.")

[22] Some philosophers currently writing on the free-will problem, following John Martin Fischer, call this thesis 'semi-compatibilism'. (That is to say, semi-compatibilism is the thesis that, *whether or not* free will is compatible with determinism, *moral responsibility* is compatible with determinism.) My only objection to that term is that attributes a rather odd sense to 'semi-': if B and C are two quite different propositions, how can the thesis that A is incompatible with B be "almost" the thesis that A is incompatible with C?