From Peter van Inwagen, *Metaphysics*, 4th edition, Westview Press, 2015.

N.B. The required reading for October 12 is just ***pp. 1-16*** *of this text. The rest will be useful later, however, when we get to libertarianism.*

Freedom of the will

We now turn to another mystery, a mystery about the powers of rational beings; that is, a mystery about what human beings are able to do. This mystery is the mystery of free will and determinism. The best way to get an intuitive grasp of the concepts of free will and determinism and the relations between them is to think of time as a "garden of forking paths," That is, to think of the alternatives one considers when one is deciding what to do as being parts of various "alternative futures" and to think of these alternative futures diagrammatically, in the way suggested by a path or a river or a road that literally forks:

Let us first consider the concept of free will. If Jane is trying to decide whether to tell all or to continue her life of deception, she is in a situation strongly analogous to that of someone who is hesitating between forks in a road. That is why this sort of diagram is so helpful to someone who is thinking about decisions and the future. To say that one has free will is to say that when one decides among forks in the road of time (or more prosaically, when one decides what to do), one is at least sometimes able to take more than one of the forks. Thus Jane, who is deciding between a fork that leads to telling all and a fork that leads to a life of continued deception, has free will (on this particular occasion) if she is able

to tell all and is also able to continue living a life of deception. One has free will if sometimes more than one of the forks in the road of time is "open" to one. One lacks free will if on every occasion on which one must make a decision, only one of the forks before one of course it will be the fork one in fact takes—is open to one. If John is locked in a room and doesn't know the door is locked, and if he is in the process of deliberating about whether to leave, one of the alternative futures he is contemplating—leaving—is in point of fact not open to him, and he thus lacks free will in the matter of staying or leaving.

It is a common opinion that free will is required by morality. Let us examine this common opinion from the perspective provided by our picture of time as a garden of forking paths. Although it is obviously false—for about six independent reasons—that the whole of morality consists in making judgments of the form 'You should not have done X', we can at least illustrate certain important features of the relation between free will and morality by examining the relation between the concept of free will and the content of such judgments. The judgment that you shouldn't have done X implies that you should have done something else instead; that you should have done something else instead implies that there was something else for you to do; that there was something else for you to do implies that you were *able* to do something else; that you were able to do something else implies that you have free will. To make a negative moral judgment about one of your acts is to evaluate your taking one of the forks in the road of time, to characterize that fork as a worse choice than at least one of the other forks open to you. (Note that if you have made a choice by taking one of the forks in what is literally a road, no one could say you should have taken one of the other forks if all the other forks were blocked.) A negative moral evaluation of what someone has done requires two or more alternative possibilities of action for that person, just as surely as a contest requires two or more contestants.

Let us now turn from the concept of free will to the concept of determinism. We shall see how thinking of time as a garden of forking paths can help us understand this concept. Determinism is the thesis that it is true at every moment that the way things then

are determines a unique future, that only *one* of the alternative futures that may exist relative to a given moment is a physically possible continuation of the state of things at that moment. Or, if you like, we may say that determinism is the thesis that only one continuation of the state of things at a given moment is consistent with the laws of nature. (For it is the laws of nature that determine what is physically possible. It is, for example, now physically possible for you to be in Chicago at noon tomorrow if and only if your being in Chicago at noon tomorrow is consistent with both the present state of things and the laws of nature.) Thus, according to determinism, although it may often seem to us that we confront a sheaf of possible futures (like this),

what we really confront is something like this:

This figure is almost shaped like a road that splits into four roads, but not quite: three of the four "branches" leading away from the "fork" are not connected with the original road, although

they come very close to it. (Thus they are not really branches in the road, and the place at which they almost touch the road is not really a fork.) If we were to view this figure from a distance—from across the room, say—it would seem to us to have the shape of a road that forks. We have to look at it closely to see that what appeared from a distance to be three "branches" are not connected with the long line or with one another. In the figure, the point at which the three unconnected lines *almost* touch the long line represents the present. The unconnected lines represent possible futures that are not *physically* possible futures—because they are not physically possible continuations of the present. The part of the long line to the right of the "present" represents a future that is a physically possible continuation of the present. The gaps between the long line and the unconnected lines represent causal discontinuities, violations of the laws of nature in a word, miracles. The reason these futures are not physically possible continuations of the present is that "getting into" any of them from the present would require a miracle. The fact that the part of the long line that lies to the right of the "present" actually proceeds from that point represents the fact that this line-segment corresponds to a physically possible future.

This figure, then, represents four futures, three of which are physically impossible and exactly one of which is physically possible. If these four futures are *all* the futures that "follow" the present, the figure represents the way each moment of time must be if the universe is deterministic: each moment must be followed by *exactly* one physically possible future.

The earlier diagram, however, represents an indeterministic situation. The road really does fork. The present is followed by four possible futures. Any one of them could, consistently with the laws of nature, evolve out of the present. Any one of them could, consistently with the laws of nature, turn out to be the actual future. It is only if the universe is indeterministic, therefore, that time *really is* a "garden of forking paths," But even in a deterministic universe, time could *look like* a garden of forking paths. Remember that our figure, when viewed from across the room, looked as if it had the shape of a road that forked. We cannot see all, or even very many, of the causes operating in any situation.

It might therefore be that the universe is deterministic, despite the fact that it sometimes seems to us human beings that there is more than one possible future. It may seem to Jane that she faces two possible futures, in one of which she tells all and in the other of which she continues her life of deception. But it may well be that the possibility of one or the other of these contemplated futures is mere appearance—an illusion, in fact. It may be that in reality, causes already at work in her brain and central nervous system and immediate environment have already "ruled out" one or the other of these futures: it may be that one or the other of them is such that it could not come to pass unless a physically impossible event, a miracle, were to happen in her brain or central nervous system or environment.

Ask yourself this question. What would happen if some supernatural agency—God, say—were to "roll history back" to some point in the past and then "let things go forward again"? Suppose the agency were to cause things to be once more just as they were at high noon, Greenwich mean solar time, on 11 March 1893 and were thereafter to let things go on of their own accord. Would history literally repeat itself? Would there be two world wars, each the same in every detail as the wars that occurred the "first time around"? Would a president of the United States called 'John F. Kennedy' be assassinated in Dallas on the date that in the new reckoning is called '22 November 1963'? Would you or at least someone exactly like you exist? If the answer to any of these questions is No, determinism is false. Equivalently, if determinism is true, the answer to all these questions is Yes. If determinism is true, then, if the universe were "rolled back" to a previous state by a miracle (and there were no further miracles), the history of the world would repeat itself. If the universe were rolled back to a previous state thousands of times, exactly the same events would follow each of these thousands of "reversions." If there are no forks in the road of time—if all the apparent forks are merely apparent, illusions due to our limited knowledge of the causes of things—, restoring the universe to some earlier condition is like moving a traveler on a road without forks back to an earlier point on that road. If there are no forks in the road, the traveler must traverse the same path a second time.

It has seemed obvious to most people who have not been exposed (perhaps 'subjected' would be a better word) to philosophy that free will and determinism are incompatible. It is almost impossible to get beginning students of philosophy to take seriously the idea that there could be such a thing as free will in a deterministic universe. Indeed, people who have not been exposed to philosophy usually understand the word 'determinism' (if they know the word at all) to stand for the thesis that there is no free will. And you might think that the incompatibility of free will and determinism deserves to seem obvious—because it is obvious. To say that we have free will is to say that more than one future is sometimes open to us. To affirm determinism is to say that every future but the actual future is physically impossible. And surely a physically impossible future can't be open to anyone, can it? If we know that a *Star Trek* sort of future is physically impossible (because, say, the "warp drives" and "transporter beams" that figure essentially in such futures are physically impossible), we know that a Star Trek future is not open to us or to our descendants.

People who are convinced by this sort of reasoning are called *incompatibilists:* they hold that free will and determinism are incompatible. As I have hinted, however, many philosophers are *compatibilists:* they hold that free will and determinism are compatible. Compatibilism has an illustrious history among English-speaking philosophers, a history embracing such figures as the seventeenth-century English philosopher Thomas Hobbes, the eighteenth-century Scottish philosopher David Hume, and the nineteenth-century English philosopher John Stuart Mill. And the majority of English-speaking philosophers in the twentieth century were compatibilists. (But compatibilism has not had many adherents on the continent of Europe. Kant, for example, called it a "wretched subterfuge.")

A modern compatibilist can be expected to reply to the line of reasoning I have just presented in some such way as follows:

> Yes, a future, in order to be open to one, does need to be physically possible—*in one sense.* I agree that a future can't be open to one if it contains faster-than-light travel and faster-than-light travel is

physically impossible. But we must distinguish between a future's being "internally" physically possible and its *having a physically possible connection with the present.* A future is internally physically possible if everything that happens in it is permitted by the laws of nature. A future has a physically possible connection with the present if it could be "joined" to the present without any violation of the laws of nature. A physically possible future that does not have a physically possible connection with the present is one that, given the present state of things, would have to be "inaugurated" by a miracle (an event that violated the laws of nature) but in which, thereafter, events proceeded in accordance with the laws. Determinism indeed says that, of all the internally physically possible futures, one and only one has a physically possible connection with the present—one and only one could be joined to the present without a violation of the laws of nature. My position is that some futures that could not be joined to the present without a violation of the laws of nature are, nevertheless, open to some people. Fortunately this does not commit me to the thesis that some of the futures open to some people are not internally physically possible—"fortunately" because that thesis is obviously false.

Two philosophical problems face the defenders of compatibilism. The easier is to provide a clear statement of *which* futures that do not have a physically possible connection with the present are "open" to us. The more difficult is to make it seem at least plausible that futures that are in this sense open to an agent really deserve to be so described.

An example of a solution to these problems may make the nature of the problems clearer. The solution I shall briefly describe would almost certainly be regarded by all present-day compatibilists as defective, although it has a respectable history. I choose it not to suggest that compatibilists can't do better, but simply because it can be described in fairly simple terms.

According to this solution, a future is open to an agent if, given that the agent chose that future, chose that path leading away from (what seemed to be) a fork in the road of time, it would come to pass. Thus it is open to me to stop writing this book and do a little dance because, if I so chose, that's what I'd do. But if Alice is locked in a prison cell, it is not open to her to leave: if she chose to leave, her choice would be ineffective because she would come up against a locked prison door. Now consider the future I said was

open to me—to stop writing and do a little dance—and suppose determinism is true. Although a choice on my part to behave in that remarkable fashion would (no doubt) be effective if it occurred, it is as a matter of fact not going to occur, and therefore, given determinism, it is determined by the present state of things and the laws of nature that such a choice is not going to occur. It is in fact determined that *nothing* is going to occur that would have the consequence that I stop writing and do a little dance. Therefore, none of the futures in which I act in that bizarre way has a physically possible connection with the present: such a future could come to pass only if it were inaugurated by an event of a sort ruled out by the present state of things and the laws of nature. And yet as we have seen, many of these futures are "open" to me in the sense of 'open' the compatibilist has proposed.

Is this a reasonable sense to give to this word? (We now take up the second problem confronting the compatibilist.) This is a very large question. The core of the compatibilist's answer is an attempt to show that the reason we are interested in open or accessible futures is that we are interested in modifying the way people behave. One important way in which we modify behavior is by rewarding behavior we like and punishing behavior we dislike. We tell people that we will put them in jail if they steal, and that they will get a tax break if they invest their money in ways we deem socially useful. But there is no point in trying to get people to act in a certain way if that way is not in some sense open to them. There is no point in telling Alfred that he will go to jail if he steals unless it is somehow open to him not to steal.

And what is the relevant sense of "open"? Just the one I have proposed, says the compatibilist. One modifies behavior by modifying the choices people make. That procedure is effective just insofar as choices are effective in producing behavior. If Alfred chooses not to steal (and remains constant in that choice), then he won't steal. But if Alfred chooses not to be subject to the force of gravity, he will nevertheless be subject to the force of gravity. Although it would no doubt be socially useful if there were some people who were not subject to the force of gravity, there is no point in threatening people with grave consequences if they do not break the bonds of gravity, for even if you managed to induce

some people to choose not to be subject to the force of gravity, their choice would not be effective. Therefore (the compatibilist concludes), it is entirely appropriate to speak of a future as "open" if it is a future that would be brought about by a choice—even if it were a choice that was determined not to occur. And if Alfred protests when you punish him for not choosing a future that was in this sense open to him, on the ground that it was determined by events that occurred before his birth that he not make the choice that would have inaugurated that future—if he protests that only a *miracle* could have inaugurated such a future—you can tell him his punishment will not be less effective in modifying his behavior (and the behavior of those who witness his punishment) on *that* account.

When things are put that way, compatibilism can look like nothing more than robust common sense. Why then do people have so much trouble believing it? Why does it arouse so much resistance? In my view, it arouses resistance because compatibilists make their doctrine look like robust common sense by sweeping a mystery under the carpet (and despite their best efforts, the bulge shows). People are aware that something is amiss with compatibilism even when they are unable to articulate their misgivings. I believe it is possible to lift the carpet and display the hidden mystery.

There are certain facts that no human being can do anything about and that no human being in history could ever have done anything about. Among these are the fact that the earth is round, the fact that magnets attract iron, the fact that there were once dinosaurs, and the fact that 317 is a prime number. Although no one would deny this it must be conceded that the concept expressed by the words "*x* can't do anything about *y* (and never could have)" is not entirely unproblematic. Consider this case. I ask you whether you can do anything about the fact that the document we need is locked in the safe. You reply, "No I can't, I don't know the combination." Or this case. Your number was not drawn in the lottery, and I ask you whether you were ever able to do anything about that (that is, whether you were ever able so to arrange matters that your number be drawn). You reply, "No, my number wasn't drawn, and I wasn't able in any way to influence

what number was drawn," Your replies certainly have a point. They would (assuming they are true statements) be excellent excuses if someone said it was wrong of you not to open the safe or maintained that you should be *punished* for failing to have a winning lottery ticket. But these facts differ in an important way from the facts in the above list of examples (the roundness of the earth and so on). You would have been able to open the safe if you had had knowledge you didn't have—or if you had made a guess about the combination and had guessed right. It could have happened that you won the lottery—*if* a different series of numbered balls had been drawn. But no knowledge, and no fantastic stroke of luck, would render you able to do anything about the shape of the earth, events in the distant past, the physical properties of iron, or the arithmetical properties of a number. Let us understand "*x* can't do anything about *y* (and never could have)" in the following very strong sense: "*x* is and always was unable to anything about *y*, and *x* would never have been able to do anything about *y*, no matter what knowledge *x* might have had and no matter how lucky *x* might have been". Even in this very strong sense of the words, it remains true that there are facts that no human being can do anything about—and that no human being in history could *ever* have done anything about: the four facts I have cited, and of course, an enormous number of others. Let us call these facts "untouchable" facts. This term is a mere label. It has no meaning beyond the meaning of the longer phrase it abbreviates. I introduce it simply to avoid having to write phrases of the form "*x* is and always was unable to anything about *y*, and *x* would never have been able to do anything about *y*, no matter what knowledge *x* might have had and no matter how lucky *x* might have been" over and over again.

The notion of an untouchable fact has a certain logic to it. One of the principles of this logic is, or so it seems, embodied in the following thesis, which I shall refer to simply as the Principle:

> Suppose it's an untouchable fact that *p.* And suppose also that the following conditional (if-then) statement expresses an untouchable fact: if *p,* then *q*.[2] It follows from these two suppositions that it's an untouchable fact that *q*.

To endorse the Principle is to endorse the following thesis: Replace the symbols '*p*' and '*q*' in the Principle with any declarative sentences you like (the same sentence must replace '*p*' at each place it occurs, and likewise with '*q*'); the result will be true. Here is an example that will illustrate what this thesis implies. We replace '*p*' with 'The last dinosaur died long before I was born' and '*q*' with 'I have never seen a living dinosaur'; the result is:

> Suppose it's an untouchable fact that the last dinosaur died long before I was born. And suppose also that the following conditional statement expresses an untouchable fact: if the last dinosaur died long before I was born, then I have never seen a living dinosaur. It follows from these two suppositions that it's an untouchable fact that I have never seen a living dinosaur.

And this statement or series of statements (the Principle tells us) is true.

Is the Principle correct? It is hard to see how anyone could deny it. How could anyone be able to do anything about something that is an inevitable consequence of something no one can do anything about? And yet as we shall see, the compatibilist must deny the Principle. To see why this is so, let us suppose that determinism is true and that the Principle is correct. Now let us consider some fact we should normally suppose was not an untouchable fact. Let us consider the fact that I am writing this book. Most people—at least most people who knew I was writing a book—would assume that this fact was not an untouchable fact because, if for no other reason, they would assume that I was (or once had been) in a position to "do something about it." They would assume that it was open to me to have undertaken some other project or no project at all. But we are supposing the truth of determinism, and that means that ten million years ago (say) there was only one physically possible future (only one physically possible continuation of the way things then were), a future that included my being engaged in writing this book at the present date (since that is what I am in fact doing): given the way things were ten million years ago and given the laws of nature, it had to come to pass that at the present time I

should be engaged in writing this book. But consider these two statements of fact

- Things were thus-and-so ten million years ago.
- If things were thus-and-so ten million years ago, then I am now writing this book.

(Here 'thus-and-so' is a sort of gesture at a complete description or specification of the way things were ten million years ago.) The facts expressed by these two statements are both untouchable facts. No human being is able, or ever has been able, to do anything about the way the world was ten million years ago. And no human being is able, or ever has been able, to do anything about the fact expressed by the second statement, for this statement is a consequence of the laws of nature, and no human being can do anything about what the laws of nature are or what their consequences are. If we imagine a possible world in which (as in the actual world) things were thus-and-so ten million years ago, but in which I decided to learn to sail instead of writing this book, we are imagining a world in which the laws of nature are different; for the *actual* laws dictate that if at some point in time things are thus-and-so, then ten million years later I (or at any rate someone just like me) shall be writing and not sailing. (Remember: we are assuming that determinism is true.)

Recall, now, the Principle. If both the above statements are statements of untouchable fact, it follows, by the Principle, that the fact that I am now writing this book is an untouchable fact. And obviously the content of the particular example—my writing a book—played no role in the derivation of this conclusion: if determinism is true and if the Principle is correct, *all* facts are untouchable facts. It follows, given the Principle, that determinism implies that there is no free will. For if anyone on any occasion has ever been able to act otherwise, then someone has been able to cause certain things to be different from the way they in fact are. And if anyone has ever had *that* ability, then some facts are not untouchable facts. This is why the compatibilist must reject the Principle. This is the hidden mystery that I contend, lies behind the facade of bluff common sense compatibilism presents to the world:

the compatibilist must reject the Principle, and the Principle seems to be true beyond all possibility of dispute. (The compatibilist who does not reject the Principle must hold that facts about what went on in the world before there were any human beings are not untouchable facts—or that facts about what the laws of nature are not untouchable facts. And these alternatives look even more implausible than a rejection of the Principle.) If the Principle were false, that would be a great mystery indeed.

We must not forget, however, that mysteries really do exist. There are principles that are commonly held, and with good reason, to be false, and whose falsity seems to be just as great a mystery as the falsity of the Principle would be. Consider, for example, the principle usually called "the Galilean Law of the Addition of Velocities." This principle is a generalization of cases like the following. Suppose an airplane is flying at a speed of 800 kilometers per hour relative to the ground; suppose that inside the aircraft a housefly is buzzing along at a speed of 30 kilometers per hour relative to the airplane in the direction of the airplane's travel: then the fly's speed relative to the ground is the sum of these two speeds, 830 kilometers per hour. According to the Special Theory of Relativity, an immensely useful and well-confirmed theory, the Galilean Law of the Addition of Velocities is not true (although what it tells us when it is applied to velocities of the magnitudes we usually consider in everyday life comes very, very close to the truth). And yet when one considers this principle in the abstract—in isolation from the considerations that guided Einstein in his development of Special Relativity—it seems to force itself up on the mind as true, to be true beyond all possibility of doubt. It seems, therefore, that the kind of "inner conviction" that sometimes moves one to say things like, "I can just see that that proposition has to be true" is not infallible. (This is not an isolated example. Consider the case of Euclidean geometry, which seems to force itself upon the mind as the real geometry of the physical world. The physicists tell us, however, that Euclidean geometry is at best *approximately true* of the physical world.)

Nevertheless, a mystery is a mystery. Since compatibilism hides a mystery, should we not therefore be incompatibilists? Unfortunately, incompatibilism also hides a mystery.

Behold, I show you a mystery.

If we are incompatibilists, we must reject either free will or determinism (or both). What happens if we reject determinism? It is a bit easier now to reject determinism than it was in the nineteenth century, when it was commonly believed, and with reason, that determinism was underwritten by physics. But the quantum-mechanical world of current physics is irreversibly indeterministic (at least this is the usual view among physicists), and physics has therefore got out of the business of underwriting determinism. Nevertheless, the physical world is filled with objects and systems that seem to be deterministic "for all practical purposes"—digital computers, for example—and many philosophers and scientists believe that a human organism is deterministic for all practical purposes. But let us not debate this question. Let us suppose for the sake of argument that human organisms display a considerable degree of indeterminism. Let us suppose in fact that each human organism is such that when the human person associated with that organism (we leave aside the question whether the person and the organism are identical) is trying to decide whether to do A or to do B, there is a physically possible future in which the organism behaves in a way appropriate to a decision to do A, and that there is also a physically possible future in which the organism behaves in a way appropriate to a decision to do B. We shall see that this supposition leads to a mystery. We shall see that the indeterminism that seems to be required by free will seems also to destroy free will.

Let us look carefully at the consequences of supposing human behavior to be undetermined. Suppose Jane is in an agony of indecision; if her deliberations go one way, she will in a moment speak the words, "John, I lied to you about Alice," and if her deliberations go the other way, she will bite her tongue and remain silent. We have supposed there to be physically possible continuations of the present in which each of these things happens. Given the whole state of the physical world at the present moment, and given the laws of nature, both these things are possible; either might equally well happen.

Each contemplated action will, of course, have antecedents in the motor speech area of Jane's cerebral cortex, for it is in that part

of Jane (or of her body) that control over her vocal apparatus resides. Let us make a fanciful assumption about these antecedents, since it will make no real difference to our argument what they are. (It will help us to focus our thoughts if we have some sort of mental picture of what goes on inside Jane at the moment of decision.) Let us suppose that a certain current-pulse is proceeding along one of the neural pathways in Jane's brain and that it is about to come to a fork, And let us suppose that if it goes to the left, she will make her confession, and that if it goes to the right, she will remain silent. And let us suppose that it is undetermined which way the pulse will go when it comes to the fork: even an omniscient being with a complete knowledge of the state of Jane's brain and a complete knowledge of the laws of physics and unlimited powers of calculation could say no more than, "The laws and the present state of her brain would allow the pulse to go either way; consequently, no prediction of what the pulse will do when it comes to the fork is possible; it might go to the left, and it might go to the right, and that's all there is to be said."

Now let us ask: Is it *up to Jane* whether the pulse goes to the left or to the right? [4] If we think about this question for a moment, we shall see that it is very hard to see how this could be up to her. Nothing in the way things are at the instant before the pulse makes its "decision" to go one way or the other makes it happen that the pulse goes one way or goes the other. If it goes to the left, *that just happens.* If it goes to the right, *that* just happens. There is no way for Jane to *influence* the pulse. There is no way for her to make it go one way rather than the other. Or at least there is no way for her to make it go one way rather than the other and leave the "choice" it makes an undetermined event. If Jane did something to *make* the pulse go to the left, then obviously its going to the left would *not* be an undetermined event. It is a plausible idea that it is up to an agent what the outcome of a process will be only if the agent is able to arrange things in a way that would make the occurrence of *this* outcome inevitable *and* able to arrange things in a way that would make the occurrence of *that* outcome inevitable. If this plausible idea is right, there would seem to be no possibility of its being up to Jane (or to anyone else) what the outcome of an *indeterministic* process would be. And it seems to follow that if,

when one is trying to decide what to do, it is truly undetermined what the outcome of one's deliberations will be, it cannot be up to one what the outcome of one's deliberations will be. It is therefore far from clear whether incompatibilism is a tenable position. The incompatibilist who believes in free will must say this: it is possible, despite the above argument, for it to be up to an agent what the outcome of an indeterministic process will be. But how is the argument to be met?

---------- *end of required reading* ----------

Some incompatibilists attempt to meet this argument by means of an appeal to a special sort of causation. Metaphysicians have disagreed about what kinds of things stand in the cause-and-effect relation. This is the orthodox, or "Humean" position: although our idioms may sometimes suggest otherwise, causes and effects are always events. We may say that "Stalin caused" the deaths of millions of people, but when we talk in this way, we are not, in the strictest sense, saying that an *individual thing* (Stalin) was the cause of certain events. It was, strictly speaking, certain *events* (certain actions of Stalin) that were the cause of certain other events (the millions of deaths). It has been suggested, however, that although events do indeed cause other events, in some cases *persons* or *agents,* individual things, cause events. According to this suggestion, it might very well be that an event in Jane's brain—a current-pulse taking the left-hand branch of a neural fork, say—had Jane as its cause. And not some event or change involving Jane, not something taking place inside Jane, not something Jane *did* but Jane herself, the person Jane, the agent Jane, the individual thing Jane.

This "type" of causation is usually labeled 'agent-causation', and it is contrasted with 'event-causation', the other "type" of causation, the kind of causation that occurs when one event causes another event. An event is a change in the intrinsic properties of an individual or a change in the ways certain individuals are related to one another. Event-causation occurs when a change that occurs at a certain time is due to a change that occurred at some earlier time. If there is such a thing as agent-causation, however, some changes

are not due to earlier changes but simply to agents: to agents *full stop*; to agents *period*. Let us now return to the question confronting the incompatibilist who believes in free will: How is it possible for it to be up to an agent what the outcome of an indeterministic process will be? Those incompatibilists who appeal to agent-causation answer this question as follows:

> "A process's having one outcome rather than one of the other outcomes it might have had is an event. For it to be up to an agent what the outcome of a process will be is for the agent to be able to cause each of the outcomes that process could have. Suppose, for example, that Jane's deciding what to do was an indeterministic process, and that this process terminated in her deciding to speak, although since it was indeterministic, the laws of nature and the way things were when the process was initiated were consistent with its terminating in her remaining silent. But suppose that Jane caused the process to terminate in her speaking, and that she *once was* able to cause it to terminate in her being silent. Then it was up to her what the outcome was. That is what it is for it to have been up to an agent whether a process would terminate in A or B: to have caused it to terminate in one of these two ways and to have been *able to* cause it to terminate in the other."

There are two "standard" objections to this sort of answer. They take the form of questions:

a. "But what does one add to the assertion that Jane decided to speak when one says she was the agent-cause of her decision to speak?"

b. "But what about the event *Jane's becoming the agent-cause of her decision to speak?* According to your position, this event occurred, and it was undetermined—for if it were determined by some earlier state of things and the laws of nature, then her decision to speak would have been determined by these same factors. Even if there is such a thing as agent-causation, and this event occurred, how could it have been *up to Jane* whether it occurred? And if Jane was the agent-cause of her decision to speak, and it was not up to her whether she was the agent-cause of her decision to speak, then it was not up to her whether she would speak or remain silent"

These two standard objections have standard replies. The first reply is "I don't know how to answer your question. But that is because causation is a mystery, and not because there is any *special* mystery about agent-causation. How would *you* answer the corresponding question about event-causation: What does one add to the assertion that two events occurred in succession when one says the earlier was the *cause* of the later?" The second is, "But it *was* up to Jane which of the two events *Jane's becoming the agent-cause of her decision to speak* and *Jane's becoming the agent-cause of her decision to remain silent* would occur. This is because she was the agent-cause of the former and was able to have been the agent-cause of the latter. In any case in which Jane is the agent-cause of an event, she is also the agent-cause of her being the agent-cause of that event, and the agent-cause of her being the agent-cause of her being the agent-cause of that event, and so on 'forever'. Of course, she is no doubt not *aware* of being the agent-cause of all these events, but the doctrine of agent-causation does not entail that agents are aware of all the events of which they are agent-causes."

Perhaps these replies are effective and perhaps not. I reproduce them because they are as I have said standard replies to standard objections. I have no clear sense of what is going on in this debate because I do not understand agent-causation. At least I don't think I understand it. To me, the suggestion that an individual thing, as opposed to a *change* in an individual thing, could be the cause of a change is a mystery. I do not intend this as an argument against the *existence* of agent-causation—of some relation between individual things and events that when it is finally comprehended, will be seen to satisfy the descriptions of "agent-causation" that have been advanced by those who claim to grasp this concept. The world is full of mysteries. And there are many phrases that seem to some to be nonsense but are in fact not nonsense at all. ("Curved space! What nonsense! Space is what things that are curved are curved *in.* Space itself can't be curved." And no doubt the phrase 'curved space' *wouldn't* mean anything in particular if it had been made up

by, say, a science-fiction writer and had no actual use in science. But the general theory of relativity does imply that it is possible for space to have a feature for which, as it turns out, those who understand the theory all regard 'curved' as an appropriate label.) I am saying only that agent-causation is a mystery and that to explain, by an appeal to agent-causation, how it could be up to someone what the outcome of an indeterministic process would be, is to explain a mystery by a mystery.

But now a disquieting possibility suggests itself. Perhaps the explanation of the fact that both compatibilism and incompatibilism seem to lead to mysteries is simply that the concept of free will is self-contradictory. Perhaps free will is, as the incompatibilists say, incompatible with determinism. But perhaps it is also incompatible with indeterminism, owing to the impossibility of its being up to an agent what the outcome of an indeterministic process will be. If free will is incompatible with both determinism and indeterminism, then since either determinism or indeterminism has to be true, free will is impossible. And of course what is impossible does not exist. Can we avoid mystery by accepting the non-existence of free will? If we are willing to say that free will does not exist, then we need not reject the Principle—and we need not suppose it is possible for it to be up to an agent what the outcome of an indeterministic process will be.

But consider. Suppose you are trying to decide what to do. And suppose the choice that confronts you is not a trivial one. Let us not suppose you are trying to decide which of two movies to see or which flavor of ice cream to order. Let us suppose the matter to be one of great importance—of great importance to *you,* at any rate. You are, perhaps, trying to decide whether to marry a certain person or whether to risk losing your job by reporting unethical conduct on the part of a superior or whether to sign a "do not resuscitate" order on behalf of a beloved relative who is critically ill. Pick one of these situations and imagine you are in it. (If you are in fact faced with a non-trivial choice, you have no need to imagine anything. Think of your own situation.) Consider the two contemplated courses of action. Hold them before your mind's eye, and let your attention pass back and forth between them. Do you

really think it isn't up to you which of these courses of action you will choose? Can you really believe that?

Some philosophers have said that although the choice between contemplated *future* courses of action always seems "open" to them at the time, when they look back on their *past* decisions, the particular decision they have made always or almost always seems inevitable from that perspective. Is this a plausible thesis? I can testify that I do not myself find any such thing when I examine my past decisions. And even if I did, I should regard it as an open question whether "foresight" or "hindsight" was more to be trusted. (Why should we suppose that hindsight is trustworthy? Maybe there is within us some psychological mechanism that produces the illusion of the inevitability of our past decisions in order to enable us more effectively to put these decisions behind us and to spare us endless retrospective agonizing over them. Maybe we have a natural tendency to interpret our past decisions in a way that presents them in the best possible light. One can think of lots of not implausible hypotheses that imply that our present impression that our past decisions were the only possible ones—if we indeed have that impression—is untrustworthy.)

When I myself look at contemplated future courses of action in the way I have described above, I discover an irresistible tendency to believe that each of them is "open" to me. This tendency may be a vehicle of illusion. It may be that free will belongs to appearance, not to reality. If the concept of free choice were self-contradictory, a belief in this self-contradictory thing might nevertheless be indispensable to human action. Let us ask ourselves: "What would it be like to believe, really to *believe,* that only one course of action is ever open to me?"

It can plausibly be argued that it would be impossible under such circumstances ever to try to decide what to do. Suppose, for example, that you are in a certain room, a room with a single door, and that this door is the only possible way out of the room. Suppose that, as you are thinking about whether to leave the room, you hear a click that may or may not have been the sound of the door's being locked. You are now in a state of uncertainty about whether the door is locked and are therefore in a state of uncertainty about whether it is possible for you to leave the room.

Can you continue to try to decide whether to leave the room? It would seem not. (Try the experiment of imagining yourself in this situation and seeing whether you can imagine yourself continuing to try to decide whether to leave.) You cannot because you no longer believe it's possible for you to leave the room. (It's not that you believe it's *im*possible for you to leave the room. You don't believe that either, for you are in a state of uncertainty about whether it is possible for you to leave.) You can of course, try to decide whether to get up and try the door. But that *is*—at least you probably believe this—possible for you. And you can try to decide, "conditionally," whether to leave the room *if* the door should prove to be unlocked. But that is not the same thing as trying to decide whether to leave the room.

This thought-experiment convinces me that I cannot try to decide whether to do A or B unless I believe that doing A and doing B are both possible for me. And therefore I am convinced that I could not try to decide what to do unless I believed that more than one course of action was sometimes open to me. And if I never tried to decide what to do, if I never deliberated, I should not be a very effective human being. In the state of nature, I should no doubt starve. In a civilized society, I should probably have to be institutionalized. Belief in one's own free will is therefore something we can hardly do without. It would seem to be an evolutionary necessity that beings like ourselves should believe in their own free will. And evolutionary necessity has scant respect for such niceties as logical consistency. It is therefore doubtful whether we can trust our conviction that we have free will (always supposing that we do have this conviction). If evolution has forced a certain belief on us (for the simple reason that we can't survive without that belief), the fact that we hold it provides no evidential support for the hypothesis that the belief is true; it does not even support the hypothesis that that belief is logically consistent. (*Aren't* there people who think that no one, themselves included, has free will? Well, there are certainly people who *say* they think this. I suspect they are not describing their own beliefs correctly. But even if there are people who think no one has free will it does not follow that these people do not think they have free will for people do have contradictory beliefs. It may be that "on one

level"—the abstract and theoretical—certain people believe free will to be an illusion, while on another level—the concrete and everyday—they believe themselves to have free will.)

Nevertheless, when all is said and done, I find myself with the belief that sometimes more than one course of action is open to me, and I cannot give it up. (Dr. Johnson has said, "Sir, we know our will is free, and there's an end on't." *I* would say, "We are *unalterably convinced* that our will is free, and there's an end on't.") And I don't find the least plausibility in the hypothesis that this belief is illusory. It can sometimes seem attractive to think of free will as an illusion. To think of free will as an illusion or to toy with the idea in a theoretical sort of way—can be attractive to someone who has betrayed a friend or achieved success by spreading vicious rumors. If you had done something of that sort, wouldn't you want to believe that you couldn't have done otherwise, that no other course of action was really open to you? Wouldn't it be tempting to suppose that your actions were determined by your genes and your upbringing or by the way things were thousands or millions of years ago? (Jean-Paul Sartre once remarked that determinism was a bottomless well of excuses.) And it is immensely attractive to suppose oneself to be a member of an intellectual elite whose members have freed themselves from an illusion to which the mass of humanity is subject. The hypothesis has its unattractive aspects too, of course. For one thing, if it rules out blame, it may well rule out praise on the same grounds. But however attractive or unattractive it may be it just seems to be false. If some unimpeachable source—God, say—were to tell me I didn't have free will, I'd have to regard that piece of information as proof that I didn't understand the World at all. It would be as if an unimpeachable source had told me that consciousness did not exist or that the physical world was an illusion or that self-contradictory statements could be true, I'd have to say, "Well, all right. You *are* an unimpeachable source. But I just don't see how what you're telling me could be true." In short: to propose that we believe that we do not have free will is to propose that we accept a mystery.

I conclude that there is no position one can take concerning free will that does not confront its adherents with mystery. I myself

prefer the following mystery: I believe that the outcome of our deliberations about what to do is undetermined and that it is nevertheless—in some way I have no shadow of an understanding of—sometimes up to us what the outcome of these deliberations will be.

I believe that if Jane has freely decided to speak, then the following must be true: if God were to create a thousand perfect duplicates of Jane as she was an instant before the decision to speak was made and were to place each one in circumstances that perfectly duplicated Jane's circumstances at that instant, some of the duplicates would choose to speak and some would choose to remain silent, and there would be no explanation whatever for the fact that any particular duplicate made whichever choice it was she made. And yet, I believe, Jane**,** the one actual Jane, was able to speak and able to remain silent. (And I believe that if all those duplicates had been created, each one, whether she spoke or remained silent, would have been able to speak and able to remain silent.)

I accept this mystery because it seems to me to be the smallest mystery available. If someone believes that human beings do not have free will that person accepts a mystery—and in my view, a greater, deeper mystery than the one I accept. Someone who denies the Principle accepts a mystery—and in my view, a greater, deeper mystery than the one I accept. Someone who denies that facts about the remote past are untouchable facts accepts a mystery—and in my view, a greater, deeper mystery than the one I accept. Someone who denies that the laws of nature are untouchable facts accepts a mystery—and in my view, a greater, deeper mystery than the one I accept. But others may judge the "sizes" of these mysteries differently,

It is important to be aware that we have not said everything there is to say about the size of the mysteries connected with the free-will problem, The most important of the topics we have not discussed in this connection is the relation between free will and morality. In our preliminary discussion of the concept of free will, we said it was a common opinion that free will was required by morality. If this common opinion is correct, then in a world without free will all moral judgments are false or in some other

way "out of place." If that were so, it would greatly aggravate the mystery confronting those who deny the existence of free will. Could it really be, for example, that racism or child abuse or genocide or serial murder are morally unobjectionable? If an unimpeachable source were to inform me that child abuse was morally unobjectionable, my dominant reaction would be one of horror. But I should also have a negative reaction to this revelation that was more intellectual, more theoretical. I should have to conclude that I didn't understand the World at all. I should have to say I simply didn't understand how it could *be* that there was nothing morally objectionable about child abuse.

It is, however, controversial whether a philosopher who rejects free will must concede that all moral judgments are false (or are all in some other way vehicles of illusion). The "common opinion" that morality requires free will is not so common as it used to be. When almost all English-speaking philosophers were compatibilists, this opinion was held by almost everyone in the English-speaking philosophical world. It was the common assumption of the compatibilists and the few incompatibilists there were. Now, however, compatibilists are less common than they used to be, owing principally to the fact that philosophers have come to realize that a compatibilist must reject the Principle. Many philosophers now reject compatibilism who might previously have been strongly attracted to this position, And because these philosophers, or many of them, believe that incompatibilism implies the impossibility of free will, they reject free will altogether. But most philosophers who reject free will are not willing to say that morality is an illusion, It has therefore become an increasingly popular position that morality does not require free will after all, For this reason, I have not included the thesis that morality is an illusion among the mysteries that must be accepted by those who reject free will. I myself continue to believe that morality is an illusion if there is no free will. (In fact, this conditional statement seems self-evident to me; if an unimpeachable source told me it was false, I'd regard its falsity as a great mystery.) But since the issues involved in the debate about this thesis pertain to moral philosophy and not to metaphysics, I will not discuss them.

However one may judge the relative “sizes” of the mysteries that confront the adherents of the various positions one might take concerning free will, these mysteries exist. The metaphysician's task is to display these mysteries. Each of us must decide, with no further help from the metaphysician, how to respond to the array of mysteries that the metaphysician has placed before us.

Suggestions for Further Reading

Berofsky’s *Free Will and Determinism* and Watson’s *Free Will* are excellent collections devoted to the problem of free will and determinism. Fischer’s more recent *Moral Responsibility* contains much useful material. My own book, *An Essay on Free Will* is a defense of incompatibilism. Large parts of it are accessible to those without formal philosophical training. The central argument of the book is attacked in Lewis’s superb article, “Are We Free to Break the Laws?” (rather difficult for those without philosophical training). Dennett’s *Elbow Room* is a highly readable (if somewhat idiosyncratic) defense of compatibilism. The question, ‘Could there be free will in an indeterministic world?’ is the main topic of the essays in O’Connor’s *Agents, Causes, and Events: Essays on Indeterminism and Free Will.*