Philosophical Perspectives, 14, Action and Freedom, 2000

LIBERTARIAN COMPATIBILISM

Kadri Vihvelin
University of Southern California

Jack was pushed by Jill, and is now tumbling down the hill. The push didn't break Jack's legs; a few minutes from now, he will get up and walk. But right now, thanks to Jill's push, Jack cannot help tumbling down the hill and thus cannot walk.

The incompatibilist thinks that there is an important sense in which there is no relevant difference between Jack's fall and the actions of any deterministic agent at any time. For example, consider deterministic Dana, who has been giving a speech to a mostly English-speaking audience and who is, at this moment, uttering an English sentence in response to a question. Dana knows how to speak Spanish, and a few minutes from now, she will say something in Spanish (when asked a question by a Spanish speaker). But right now, thanks to the deterministic causes of her English utterance, Dana *cannot* help speaking English and thus cannot say something in Spanish.

The compatibilist disagrees, and points to the differences between Jack and Dana to explain why. Jack's falling is not a voluntary action; even if he chose to walk, he would still fall. But Dana's speaking English is a voluntary action, under her volitional control; if she chose to speak Spanish instead, she would do so.

True enough, replies the incompatibilist, but the relevant question is whether Dana *can*, in the circumstances that in fact obtain, *choose* to speak Spanish. To the incompatibilist, it seems obvious that the deterministic causes of Dana's choosing to speak English are in all relevant respects like the deterministic causes of Jack's falling down the hill; they render Dana powerless to choose to speak in Spanish, or, for that matter, to make any choice other than the choice she actually makes.

Here, where argument should start, is where it usually ends. Instead of defending the assumptions about causation and laws of nature which underlie the claim that deterministic causes are, in all relevant respects, like pushes and shoves, libertarians and other incompatibilists appeal to "intuitions" about the "fixity" of the past and the laws of nature.[1] And instead of addressing seriously

these modal and metaphysical concerns, the standard compatibilist response is still some variation on the "we're the only game in town" strategy, which consists chiefly of arguing that the libertarian conception of freedom is metaphysically dubious, perhaps downright incoherent, and, at any rate, not the kind of freedom worth wanting.[2]

The only way to get past this impasse is to take seriously the legitimate incompatibilist worry that Dana is as unable to do otherwise as Jack is unable to avoid falling downhill. And the only way to do this is to understand how our ordinary ways of thinking about what we can and cannot do make libertarianism such a natural and appealing view. I think that once we've done this, we'll see that what has gone wrong in the traditional debate is that compatibilists and incompatibilists have been talking past each other. Compatibilists should be understood as arguing that determinism doesn't rule out the possession of abilities, including the abilities that have traditionally been thought necessary for free will and moral responsibility. But incompatibilists are right to insist that free will also requires the genuine opportunity of doing something other than what one in fact does. There is no incoherence or mystery in an account of free will that combines both elements. It may look as though this account is committed to incompatibilism, but I will argue that this is not the case. Just as we may have the freedom compatibilists want regardless of whether determinism is true or false, so too we may have the freedom libertarians want regardless of whether determinism is true or false.

All this will take some arguing. Before I begin, let's take a quick look at the current state of the art in the debate between incompatibilists and compatibilists.

The Incredible Ability Argument for Incompatibilism

There is really just one *argument* in the current literature in support of incompatibilism.[3] It's a reductio that goes like this:

Suppose that determinism is true. Then for every action X that I perform, there is a true historical proposition H about the intrinsic state of the world at some time prior to my birth and a true proposition L specifying the laws of nature, such that H and L jointly imply that I do X. Now let's suppose that I nevertheless have the ability to do otherwise. If so, then I have the ability to do something such that if I did it, then either H or L, or both, would be false. And if that's so, then I have either the ability to change the past or the ability to change the laws or, perhaps, the ability to do both. But to suppose that I have either of these abilities is incredible. Therefore, it cannot be the case both that determinism is true and that I have the ability to do otherwise.

This argument—let's call it the 'Incredible Ability' argument—doesn't work. The standard compatibilist reply[4] is to distinguish between two counterfactuals:

(C1) If S had done otherwise, the past would have been different.
(C2) If S had done otherwise, this would *have caused* the past to have been different.

Having distinguished **C1** from **C2**, the compatibilist points out that there is a corresponding ambiguity between two ability claims:

(A1) S has the ability to do something such that if she did it, the past would have been different.

(A2) S has the ability to do something such that if she did it, this would *have caused* the past to have been different.

The problem with the Incredible Ability argument is that it equivocates between these two ability claims. To count as a reductio against the compatibilist, the argument must establish that the compatibilist is committed to **A2**. But the compatibilist is committed only to **C1** and hence only to **A1**. The compatibilist is committed only to saying that deterministic agents have abilities which they would exercise *only if* the past had been different in the appropriate ways and there is nothing incredible about this. Consider the ability to shoot and wound a human being. Joe's got the ability, but he would exercise it only in a narrow range of circumstances—self-defense or defense of his family. As a matter of fact, these circumstances never arise and Joe never shoots anyone. But he's still got the ability.

The incompatibilist needs a new argument. But when we look to the literature, there is not much argument to be found, as opposed to appeal to "intuitions" about our powerlessness with respect to the past and the laws of nature.

Ability, Opportunity, and the Beginnings of a New Argument for Incompatibilism

Compatibilists argue that in our everyday sense of 'free', 'can', or 'is able', we do not take the claim that someone is able to walk, talk, or play the piano as entailing that the person can do these things, given the circumstances that in fact obtain. Incompatibilists insist that if we are asking what someone is able to do, then we must be asking what that person can do *in the circumstances that in fact obtain*; that is, in asking what that person can do, *given the actual past and the laws*. In the literature on free will and determinism, this debate often seems to come down to an irresoluble dispute about what we *ought* to mean by words like 'free', 'can', 'power', and 'ability'.

I think we can make progress by recasting the free will/determinism debate in terms of two concepts that are part of our ordinary thinking about what we can and cannot do.

Commonsense recognizes a distinction between abilities (understood as skills, capacities, or knowing how) and opportunities. Someone may have the ability to do something (e.g., play the piano) but be unable to do it because she lacks the opportunity (there is no piano handy). Someone else may have the opportunity (the piano's right in front of her) but be unable to play because she lacks the ability (she never took lessons). And another person may be unable to play because she lacks both ability and opportunity.

Facts about ability and opportunity are relevant to questions of freedom and responsibility in at least two different ways.

First, we ordinarily think that someone is morally responsible for failing to do something (as opposed to merely failing to *try* to do something) only if she had the opportunity as well as the ability to do it. We don't hold someone responsible for failing to play the piano if she doesn't know how to play. But we also don't hold her responsible if any of the following are true: there is no piano handy; there's a piano but it's not working; there's a piano but the person is in chains or otherwise prevented from reaching the piano.[5]

Second, we think that the range of alternatives among which we have reason to deliberate is limited to those acts that we in some sense *can* do, and we think that the relevant sense of 'can' entails both ability and opportunity. If there's no way for you to rescue a drowning child, then there's no point in deliberating about how to do so. But there are two different ways in which you may be unable to save the child. You may be unable because you lack a relevant ability—the only way to reach the child is by swimming and you cannot swim. Or you may be unable because you lack opportunity, e.g. if any of the following are true: the only way to reach the child is by boat and there is no boat; by the time you notice the child, it's already too late to save her; there are sharks in the water which will eat you before you get to the child.

We may use this distinction between ability and opportunity to help us get more clear about why determinism is supposed to be incompatible with free will and moral responsibility. Is it because determinism robs us of opportunities or is it because determinism deprives us of abilities?

To answer this, we have to say a bit more about what it is to have an ability. We make judgments about ability on the basis of evidence of a reliable causal correlation between someone's attempts to do a certain kind of act and the success of her attempts. There is a continuum of cases, ranging from the person with no ability to do X who keeps trying and sometimes gets lucky, to the person just learning to do X who succeeds more often, to the person highly skilled at X-ing, who typically succeeds most often. But since success depends partly on circumstances outside the person's control, this correlation is evidence of ability, not constitutive of ability. What, then, is it to have the ability to do X?

Here's a sketch of an account: Someone has the ability to do X just in case it's true that there are some reasonably specifiable circumstances C (e.g., working piano nearby, the person is not bound or otherwise physically prevented from reaching the piano) such that if she tried, in circumstances C, to do X, she would probably succeed.[6] We evaluate this counterfactual by considering possible worlds where our laws obtain, where the person is as similar to the way she actually is as is compatible with her trying to do X, and where circumstances C obtain. If all or some reasonably high percentage of these worlds are worlds at which the person's attempt succeeds, then she has the ability to do X.

There are cases where it's not clear whether someone is unable to do something because she lacks the ability or because something prevents her from ex-

ercising the ability she continues to possess. Consider, for instance, the pianist who suffers from stage-fright so extreme that her hands begin to shake when she tries to play in front of an audience. Should we say that her stage-fright prevents her from exercising her ability to play? Or should we say that her stage-fright temporarily deprives her of the ability to play? Or should we say that she has one ability (the ability to play while no one's watching) while lacking another ability (the ability to play while someone's watching)? How we answer these questions depends on two variables: what we think the relevant ability is, and what we choose to include in the circumstances C in terms of which the ability is defined. There is room to argue about these kinds of cases. But I think that everyone should agree that the having of abilities, understood as skills or knowhow, is compatible with the truth of determinism. After all, abilities are defined in terms of conditionals about what would probably be the case, given our laws and the relevant enabling circumstances C, *if* someone tried to do an act of the relevant kind. Such conditionals may be true even if it's also true (as the incompatibilist thinks) that no person can ever *try* to do anything that she doesn't in fact try to do. Given this, we should not think that determinism deprives us of freedom by depriving us of abilities.

Note that everything I've just said applies to "inner"or mental abilities as well as abilities to perform bodily actions like piano-playing and swimming. A deterministic agent may have the kinds of mental abilities that have traditionally been thought necessary for free will—the ability to reason and deliberate concerning possible actions, the ability to make decisions about what to do on the basis of prudential and moral considerations, and so on. Indeed, for any *ability* you might think relevant to the question of free will and moral responsibility, there is no reason to think that deterministic causal laws would deprive us of this ability.[7]

I think that the disagreement between compatibilists and incompatibilists is best understood as a disagreement about whether the deterministic causes of an action prevent a person from doing anything else, including those things she has the ability to do. That is, the incompatibilist thesis should be understood as the claim that determinism deprives persons of the *opportunity* to do anything other than what they in fact do.

Remember Jack, who was pushed by Jill, and who is now tumbling down the hill. Jack still has the ability to walk because in the relevant circumstances C (standing on solid ground, not chained or otherwise prevented from moving his limbs) if he tried to walk, he would probably succeed. But Jill's push has rendered him temporarily unable to exercise his ability. He still has the ability, but lacks the opportunity.

And remember deterministic Dana, who is at this moment answering a question in English. No one denies that Dana has the ability to speak Spanish or the ability to deliberate concerning the pro's and con's of speaking in English versus speaking in Spanish. The only sensible question is whether the deterministic causes of Dana's English utterance temporarily prevent her from exercising her ability to speak Spanish. That is, the question at issue is (or

should be) whether deterministic causes deprive Dana of the *opportunity* to speak Spanish.

If this is right, then we should understand the free will debate as a debate about whether deterministic causes are like pushes and shoves, temporarily depriving us of the opportunity to exercise our unexercised abilities. And once we understand this, incompatibilist intuitions start to make a lot more sense.

Although Jack has the ability to walk, he can't walk right now. Why not? Well, if he had walked, it *would have to have been the case that the past was different*—that Jill didn't push him. Given that she did push him, he cannot walk—he lacks the opportunity. Suppose that Dana is a deterministic robot, designed to respond in English to English questions and in Spanish to Spanish questions. Dana has the ability to speak in Spanish because there are relevant circumstances C (e.g.. she's asked a question in Spanish, she is not gagged or otherwise physically prevented from speaking) in which it's true that if she tried to speak in Spanish, she would probably succeed. But if Dana had spoken in Spanish right now, *it would have to have been the case that the past was different*—that the question to which she is responding was a Spanish question. Given that the question was in fact an English question, she cannot reply in Spanish—she lacks the opportunity.

Compare Dana to indeterministic Ingrid, who has also been giving a speech in English and who is also uttering an English sentence in response to a question. Like Dana, Ingrid has the ability to speak Spanish. But it's *false* that if she had spoken in Spanish right now, *the past would have to have been different*. If she had spoken in Spanish, the past would still have been exactly the same. Well, perhaps that's going too far. She's got no reason to speak in Spanish, so perhaps it's true that if she had spoken in Spanish, this might have been because something about the past was different (e.g.. she was asked a question in Spanish). But the past would not *have to have been different*. It might have been exactly the same—Ingrid might simply have decided to utter a Spanish sentence, perhaps as a philosophical example, perhaps as a private joke.

This apparent counterfactual difference between indeterministic Ingrid and deterministic Dana, is, I think, what fuels the incompatibilist intuition that there is a real and interesting difference between deterministic and indeterministic agents. And the counterfactual true of Ingrid—that if she did otherwise, the past would or at least *might* still have been exactly the same—is, I think, what lies behind the otherwise obscure thought that what's necessary for free will is the "categorical" or "unconditional" power to do otherwise—the power to do otherwise given *all the facts*, or, at least, the power to do otherwise *given all the facts about the past*. Let's give a name to this somewhat obscure thought and understand it as follows:

> **Fixed Past Assumption (FPA):** A person has free will only if it is at least sometimes true[8] that she has the *opportunity* as well as the ability to do otherwise; that is, only if it's at least sometimes true both that she has the ability to do something else X and also true that *if she had tried and suc-*

ceeded in doing X, the past prior to her choice would or at least might still have been exactly the same.[9]

If **FPA** is true, then it looks as though there is a fairly straightforward argument for the thesis that free will is possible, if possible at all, only at indeterministic worlds, that is, only at those worlds where more than one future course of events is nomologically possible, given the same past. But should we accept **FPA**?

FPA has this much going for it. It is part of a deeply engrained picture we have of ourselves as agents, a picture which represents our relation to the past as fundamentally different from our relation to the future. The past is "fixed", over and done with; when we make decisions we must assume that the past is the way it is. The future, on the other hand, is not yet fixed, it is in some sense "open", up to us, under our control. By making a choice (or reaching a decision or forming an intention) and then acting on it, we actualize the future; it is *we* who are responsible for the future being the way it is.

But the fact that a picture is intuitively appealing doesn't mean that it's correct, or even that it can survive closer scrutiny.

Here's a first attempt at an argument in defense of **FPA**: We are trying to give an account of the facts that make it true that *nothing* prevents someone from exercising her ability to do X. If it's really true that nothing prevents S from exercising her ability to do X, it must be true that S can do X, *given all the circumstances which in fact obtain*. But that's just shorthand for "S can, given *all the facts*, do X". If we ask whether the imprisoned lifeguard can save the drowning child, we understand this as a question about what she can do, given all the facts, including the fact that the door was locked a few minutes ago. We aren't asking the *different* question of what she might have been able to do if the past had been different.

But this is not a good argument. *All* the facts include facts about the *future*. For any action X that anyone fails to do, the totality of facts includes the fact that she will *not* do X. If we insist on saying that someone can do X only if she can do X, given *all* the facts, then no one can ever do otherwise, regardless of whether determinism is true or false.

We can give a better argument in defense of **FPA**. It goes like this: The difference between past and future is not ontological; it is relational. The difference is that *we* can *cause* future events, but not past events. Of course the future would be different—would have to be different—if Jill had not pushed Jack. But that's because Jill would have *caused* it to be different, and so doesn't count against Jill's opportunity to do otherwise. If Jill had done otherwise—if she had not pushed Jack—everything might still have been just the same except for Jill's choice and its causal consequences. But if Jack had done otherwise—if he had walked—the world would have to have been different in ways that he would not have caused. If he had walked, it would have to have been the case that he isn't tumbling down the hill and for this to be true, it would have to have been the case that Jill didn't push him.

We may now formulate the claim that underlies **FPA**:

> **Agent Causation Assumption (ACA):** A person has free will only if it's at least sometimes true that she has the *opportunity* as well as the ability to do otherwise; that is, only if it's at least sometimes true both that she has the ability to do something else X and also true *that if she had tried and succeeded in doing X, everything except her choice, action, and the causal consequences of her choice and action would or at least might have been just the same*.[10]

The assumption being made here is the following: Someone has the opportunity to exercise her ability to do X just in case there is no impediment to her doing X. If there is no impediment to her doing X, then *nothing would have to be different* in order for it to be true that her attempt to do X succeeds—except, of course, her choice, action, and the causal consequences of her choice and action.

This assumption seems reasonable. Consider another kind of case, a case that has traditionally been regarded as a counterexample to compatibilist attempts to provide a conditional analysis of 'could have done X'. Mary, a native English speaker, is under general anesthetic while she undergoes surgery. According to a standard conditional analysis of 'can do X', it's true that Mary can speak English just in case it's true that if Mary chose (or tried, decided, intended, etc.) to speak English, she would succeed in speaking English. Since Mary's choosing (trying, etc.) to speak English requires her to be conscious, the conditional 'if Mary chose (tried, etc) to speak English, she would succeed' is true, so according to the conditional analysis of 'can', Mary can speak English. But since Mary is *in fact* unconscious, she cannot speak English.

Incompatibilists use cases like that of Mary to argue that 'can' is "categorical" rather than "conditional". I would draw a somewhat different moral. I think that this kind of case[11] illustrates how compatibilists and incompatibilists so often talk past each other. There is a sense in which Mary can speak English, but there is also an important and relevant sense in which she cannot. In one sense, the ability sense, Mary *can* speak English. It continues to be true of Mary, at all times while she remains unconscious, that *if* she tried, in the appropriate enabling circumstances C, to speak English, she would probably succeed. Her state of unconsciousness no more deprives her of the *ability* to speak English than Jack's state of tumbling downhill deprives him of the ability to walk. On the other hand, there is a significant sense in which Mary *cannot* speak English. Her state of unconsciousness prevents her from choosing, deciding, intending or in any way trying to bring it about that she speaks English and thus prevents her from exercising her ability. So long as she remains unconscious, Mary retains the ability to speak English, but lacks the opportunity.

Mary's case isn't a counterexample to the compatibilist thesis that *abilities* are correctly analysed in terms of conditionals. But the fact that she can't, *in the relevant sense*, speak English during the time that she's under general an-

esthetic shows that there is more to our ordinary sense of 'can' (or 'free', or 'is able') than simply having the ability to do something. And it shows that we can meaningfully speak of someone being unable to exercise an ability despite the fact that there are no external impediments to her doing so. It thus supports the incompatibilist's thesis that a person may be deprived, by conditions "beneath the skin", of the opportunity to exercise any of her unexercised abilities. And it also supports **ACA**, which explains Mary's lack of opportunity in terms of a different kind of conditional: If Mary had succeeded in speaking English, there would have to have been a difference not caused by her choice or action; she would have to have been conscious.

With these distinctions in hand, we are now ready to formulate a new argument for incompatibilism.

The No Opportunity Argument for Incompatibilism

1. Someone S has free will only if it's at least sometimes true that she is able to do otherwise.
2. S is able to do otherwise iff she has the ability to do something else X and the opportunity to do X (that is, iff she has the ability to do X and nothing prevents her from exercising the ability).
3. For any X such that S has the ability to do X, S also has the opportunity to do X only if it's true that if S had tried and succeeded in doing X, everything except her choice, action, and the causal consequences of her choice and action would or at least might have been just the same.
4. If determinism is true, then for any X such that S does not do X, if S had done X, neither her choice nor her action would have *caused* the past (prior to her choice) to be different.
5. If determinism is true, then for any X such that S does not do X, if S had done X, the past prior to S's choice would have been different.
6. Therefore, if determinism is true, for any X such that S has the ability to do X but S does not do X, S lacks the opportunity to do X.
7. Therefore, if determinism is true, S is never able to do otherwise.
8. Therefore if determinism is true, S lacks free will.

The first two premises are (or should be) common ground between the compatibilist and the incompatibilist. The third premise is entailed by **ACA**. The fourth premise relies on the uncontroversial assumption that there is no backwards causation, neither at the actual world nor at the closest worlds where determinism is true. The fifth premise is an assumption about counterfactuals at deterministic worlds which seems plausible.[12] Premise 6 follows from 3, 4, and 5. Premise 7 follows from 2 and 6. The conclusion follows from premises 1 and 7.

The compatibilist can't object to this argument on the grounds that it's based on a conception of free will which is incoherent, unsatisfiable, or requires a mysterious, nonnaturalistic conception of our abilities. The argument does not

claim that indeterministic agents have *abilities* lacked by deterministic agents and is therefore compatible with a naturalistic understanding of abilities as complex dispositional properties. The argument is neutral between different accounts of what abilities constitute free will; it insists only that a *necessary* condition of free will is that it is at least sometimes true that the person has the opportunity as well as the ability to do otherwise. Finally, the opportunity component of free will, understood in terms of **ACA**, is neither mysterious nor logically unsatisfiable—indeterministic Ingrid satisfies it.

The problem for the compatibilist *seems* to be that she has to reject premise 3 and therefore **ACA**. But it isn't clear that there are good reasons for rejecting **ACA**. **ACA** says that a necessary condition for having free will is the truth of a counterfactual that is, I have argued, intuitively relevant to our beliefs about what we have the opportunity to do. The compatibilist cannot insist that facts about opportunity are facts "outside the skin" (e.g.. facts about the absence of locked doors and chains). For nothing turns on our use of the word 'opportunity'. The incompatibilist's worry about deterministic causes is that they are the internal equivalent of pushes and shoves, temporarily preventing the agent from exercising most of the abilities she continues to possess. I have argued that **ACA** is the best way of making precise this worry.

Given this way of understanding the incompatibilist's argument, it is hard to resist the conclusion that the freedom to do otherwise is incompatible with determinism.[13] And some compatibilists have not resisted. An increasing number of compatibilists have embraced the view that John Fischer has christened "Semi-Compatibilism"; they grant that free will (understood as entailing the freedom to do otherwise) is incompatible with determinism, but insist, following Harry Frankfurt, that moral responsibility is nevertheless compatible with determinism.[14]

Agency and Counterfactuals

But it's not clear that the compatibilist *has* to reject **ACA** and premise 3. The No Opportunity argument relies on a thesis about counterfactuals—premise 5. For premise 5 to be true, it must be the case that, for every deterministic agent S, and every relevant context of utterance, the sentence below expresses a true proposition:

> **Deterministic Backtracker (DB):** If S had done otherwise, the past prior to her choice would have been different.

If, on the other hand, there is a relevant context at which **DB** fails to express a true proposition, then premise 5 is false and the No Opportunity argument fails.

I will argue that we should reject the claim that **DB** is always true. I will argue for this in two stages. First, I will make some observations about how we

in fact evaluate counterfactuals in certain contexts. Then I will argue that we are justified in evaluating the relevant counterfactuals this way and that our justification is independent of the truth or falsity of determinism.

Consider the kinds of counterfactuals we take seriously in contexts of deliberation or decision-making *before* action and in contexts where someone is called upon to defend her action *after* the time of action. Let's call these "agency counterfactuals". Here's an example: Sara is deliberating at noon about whether or not to step on the ice after a warm winter morning during which the ice has mostly melted. Sara is a sensible and cautious person who would never step on ice unless she were sure that it's strong enough to support her weight. Knowing this fact about Sara's character, we might try to lure her onto the ice with this argument: "If you stepped on the ice, it would not have melted and it would be strong enough to support your weight." Sara would not take this argument seriously. She would agree that her character is such that we have good reason for believing that if she steps on ice, it's safe, and thus good reasons for believing that if she steps on ice, it didn't melt shortly before she stepped on it. She might even agree that there is a way of understanding the "backtracking" counterfactual "if Sara stepped on the ice, it would not have melted" which makes it true. But she believes that the counterfactual *relevant to her decision* is this one: "If I step on the ice, it would still have melted this morning, and it would not be strong enough to support my weight." And when we later asked her why she didn't walk on the ice, she replies: " Because if I had stepped on it, I would have fallen through." [15]

In considering whether or not to step on the ice, Sara assumes that, regardless of what she chooses, the past prior to her choice would still have been the way it in fact is—the ice would still have melted, she would still weigh what she actually weighs, she would still not have been supported by a helium-filled balloon, and so on. When thinking this way, Sara rejects **DB** and assumes that the following counterfactual is both true and relevant to her decision-making:

> **Fixed Past Counterfactual (prospective):** If I stepped on the ice now, at noon, the past prior to my choice would still have been just the same.

Sara's rejection of **DB** isn't restricted to the context of deliberation prior to action. When we ask her the next day why she declined to take advantage of the opportunity to walk across the ice, her reply suggests that she believes that the following counterfactual is both true and relevant to the justification of her action:

> **Fixed Past Counterfactual (retrospective):** If I had stepped on the ice just then, at noon, the past prior to my choice would still have been just the same.

Here's another fact about Sara: She believes that agency counterfactuals have truth-conditions that are objective in the following sense; she believes that

they are true or false in virtue of facts that are independent of her reasons for believing them. These facts include facts about the past. She believes that if she relies on a mistaken belief about the past in believing a counterfactual, then her counterfactual belief may turn out to be mistaken, even though it was rational for her to believe it, given everything she knew at the time. For instance, suppose that she is standing, on a different day, with her friend Stan at the edge of the ice having an argument about whether the ice would support their weight. She says: " It's safe; I went skating yesterday." He answers: "Yes, but there was a thaw this morning." They don't go on the ice, but a few minutes later they watch Jill performing the experiment and falling through. Sara says: "You were right and I was wrong; I thought the thaw could not have melted all that ice, but I see now that it did. If I had stepped on it, I would have fallen through."

I claim that Sara is representative; this is how we in fact evaluate agency counterfactuals. This is an empirical claim. My evidence consists in the following. Despite the fact that the sentences used to assert or entertain counterfactuals are highly context-sensitive and often ambiguous between different readings[16], there are some counterfactuals which we so uncontroversially accept as true that the philosophical problem (as Goodman first pointed out[17]) is to give a theory of counterfactuals which explains how this knowledge is possible. Goodman was primarily interested in the singular causal counterfactuals associated with laws[18], but everything he said applies also to that species of singular causal counterfactual I've been calling 'agency counterfactuals'. Somehow we know how to go about determining whether these counterfactuals are true or false. How do we do it? My claim, in a nutshell, is that our knowledge of the truth-conditions for agency counterfactuals is best explained by our implicit acceptance of a theory which tells us to evaluate these counterfactuals by using everything we know about the past until just before the agent's choice[19], and then reasoning from the choice onwards in accordance with what we know about the laws and causal generalizations. And our knowledge of the truth-conditions of other singular causal counterfactuals suggests that we do so by using a theory which is an extension of our theory of agency counterfactuals; we hold the past constant until either the time of the antecedent or as near to it as is compatible with the occurrence of a local "divergence miracle"[20] and then reason from there in accordance with what we know about the laws and causal generalizations. In other words, we evaluate singular causal counterfactuals in the way that David Lewis tells us we should.[21] And if determinism is true, then, whether we realize this or not, we evaluate agency counterfactuals by considering worlds where the *agent's choice* is the divergence miracle.[22]

I think, moreover, that we are justified in evaluating agency counterfactuals this way. We are justified because the theory which tells us to evaluate them this way is the theory that best accounts for the data—our knowledge of the counterfactuals we uncontroversially accept as true.

A brief historical digression will help to defend my claim. In his seminal article, Goodman drew our attention to something he called 'the problem of cotenability'. Consider a dry well-made match which in fact is not struck. In the absence of specific reasons for believing otherwise (eg. someone lurking by, ready to pour water on the match at the first sign of an attempt to light it), we believe that if the match had been struck, it would have lit. But if, as Goodman assumed, the truth-conditions of a counterfactual are given by the fact that there is a valid argument from the antecedent and some true premises to the consequent, then why do we believe *this* counterfactual about the match rather than any of the other counterfactuals whose consequents may also be deduced from the antecedent, the laws and other true premises—eg. if the match had been struck, it would have been wet; if the match had been struck, there would have been no oxygen? In other words, why do we regard the dryness of the match and the fact that there is oxygen as "cotenable" premises while rejecting its unlit state as a cotenable premise? More generally, on what grounds do we choose, from among *all the truths* about the world, which ones are eligible to be counted as the premises for the purpose of constructing a valid argument to evaluate a counterfactual? Goodman named this "the problem of cotenability" and he thought it fatal to the prospects of giving a noncircular account of the truth-conditions of counterfactuals. For it seems that our grounds for accepting the dryness of the match and the presence of oxygen as cotenable premises are our beliefs in other *counterfactuals*: eg. our belief that if the match had been struck, it would still be dry and our belief that if the match had been struck, there would still be oxygen.

Goodman wrote in 1947, and what philosophers think about counterfactuals has changed considerably since then. Some philosophers have argued that what Goodman showed us is that *no* counterfactuals have objective truth-conditions; whether a counterfactual is true or false (or, on some views, "assertible" or not) is *always* determined by the facts the speaker chooses to hold fixed. On this view, there are no empirical grounds for choosing between "if the match had been struck, it would have lit" and "if the match had been struck, it would have been wet". (Or between "if Sara had stepped on the ice, she would have fallen through" and "if Sara had stepped on the ice, it would have been safe".) Other philosophers have drawn the opposite conclusion; they have taken the moral to be that our knowledge of causal truths—both actual and counterfactual—outruns our knowledge of laws. Somehow we know that striking that match *would cause* it to light, whereas striking the match would *not cause* it to get wet and wouldn't cause it not to be in the presence of oxygen even though, so far as the *laws* are concerned, all that's ruled out is that all of the following facts obtain: the match is dry; it's wellmade; it's in the presence of oxygen; it's struck; it doesn't light.

I think the first view—the rejection of objective truth-conditions for counterfactuals—is the counsel of despair, not justified at this relatively early stage of the game in our understanding of counterfactuals. After all, the first big ad-

vance in understanding counterfactuals did not come until the advent of possible worlds semantics in the early 1970's.[23] This isn't a solution to Goodman's problem but it does provides a helpful neutral framework for discussing the problem. On the possible worlds approach, the counterfactual "if P, it would be the case that Q" is true just in case the closest worlds where P is true are all worlds where Q is also true.[24] The problem of cotenability then becomes the problem of providing an account of the factors that determine which worlds count as the closest for the purpose of evaluating the counterfactual. The philosophers who reject objective truth-conditions for all counterfactuals are in effect saying that there is *never* a standard way of resolving the vagueness and ambiguity of counterfactual sentences; 'closest' *always* means 'most like the actual world in the ways that matter to the speaker'. Note that if this view is right, then the No Opportunity argument fails. For the No Opportunity argument relies on the claim that **DB** is always true. If counterfactuals lack objective truth-conditions, then we may reject **DB** by simply choosing to evaluate agency counterfactuals by holding the past fixed.

The second approach is, I think, the correct one, but we need to say more to turn it into a theory. It's not enough to gesture in the direction of "a causal theory of counterfactuals"; we need an account that tells us *which* causal facts are the ones that determine which worlds count as closest for the purpose of evaluating the counterfactual "if E had happened, then F would have happened".[25] For as a matter of fact, E *didn't* happen, so the causal facts include facts about the causes of the non-occurrence of E, facts about the effects of the non-occurrence of E, as well as facts about causal regularities. Any world where E happens will differ from the actual world with respect to some of these causal facts. So until we have an account which tells us which causal facts are the relevant ones, we don't have a solution to Goodman's problem.

Here's a first thought. (For reasons that will become obvious in a moment, let's call it 'the Naive theory'.) The relevant causal facts are facts that *we lack the power to causally affect*. We lack the power to causally affect either the *laws* or the *past*, so we hold these facts constant when we evaluate a counterfactual. So, for instance, we evaluate 'if the match had been struck, it would have lit' by considering worlds where the laws are exactly the same and the past is exactly the same until just before the time of the antecedent, when some person picks up the (dry, wellmade, unlit, in the presence of oxygen) match and strikes it. This *seems* to solve Goodman's problem in a way that accounts for our knowledge of the relevant counterfactual. For our knowledge includes, not just knowledge of regularities and laws, but also knowledge of our past interactions with the world, including our attempts, both successful and unsuccessful, to manipulate objects to bring about results we want. And our experience tells us that by striking a dry wellmade, etc. match we often succeed in lighting it, but don't (barring unusual circumstances) succeed in bringing it about that the match is wet or deprived of oxygen or not well-made.

There is something to this idea. Our beliefs about counterfactuals, causation, and our causal powers as agents are closely linked. Some philosophers have argued that we would not have the concept of causation if we didn't believe that we are agents with the power to manipulate objects and to originate causal chains.[26] Other philosophers have argued that our beliefs about agency and fixed past counterfactuals are so deeply rooted in our way of thinking of ourselves and the world that, even though "backwards" causation is not a logical impossibility, we would not accept any evidence as showing us that an *agent* has the power to causally affect the past.[27] Nevertheless, the Naive theory cannot be right, for at least two reasons. First, not all causal counterfactuals are about causes that may, even in principle, be brought about by the intervention of an agent. Second, the truth of determinism doesn't rule out either the truth of singular causal counterfactuals or our knowledge of them. But if determinism is true, there are *no worlds* with our laws and the same past where anyone chooses otherwise, thereby making something different happen. A theory of singular causal counterfactuals should apply to causes that are not manipulable by agents as well as those which are, and it should apply to deterministic as well as indeterministic worlds. So our theory of these counterfactuals should not *insist* that both the laws and the past be kept constant.

Since the counterfactuals about which we seem to be most confident are agency counterfactuals, let's begin by trying to give an account of these counterfactuals that will apply to deterministic as well as indeterministic worlds. (Perhaps we can later find a way of applying or extending the account to other singular causal counterfactuals, but we won't start by assuming we can do this.) If determinism is true, there are *no* worlds with exactly the same laws and exactly the same past where anyone chooses and does otherwise, so we must decide which respect of similarity matters more. Let's begin historically and assume it's the laws. This, of course, doesn't solve Goodman's problem, since what gave rise to that problem is the fact that the truth about the laws underdetermines the true counterfactuals. But perhaps we can use our general causal knowledge to help us out. We know that causes are temporally prior to their effects and we know that there is no direct causation at a temporal distance; the past causes the future only by way of the present. These two facts about causation suggest the following theory, which we will call "the Fixed Law" theory.[28]

The Fixed Law theory says that the closest worlds where an agent S does something X at time t are worlds with the *same laws* as the actual world and which are otherwise as *similar to the actual world at time t* as is compatible with S doing X. If determinism is true, then these worlds have the same deterministic laws as ours and thus a different causal history leading up to S's doing X. But since the past causes the future only by way of the present, we can ignore these differences, focusing on the relevant facts at time t. So, for instance, we evaluate "if S had struck the match at t... ." by considering worlds which *at time t* are very much like the actual world in all the intuitively rele-

vant respects (the match is dry, wellmade, in oxygen, unlit) except for the fact that the match is in S's hand, being struck. Since these worlds have our laws, they are all worlds at which the match lights a moment later.

This theory *seems* able to accommodate the agency counterfactual about Sara and the ice. Granted, there are worlds with our laws where Sara steps on the ice without falling through. But, as Jonathan Bennett points out[29], these are worlds which are more like ours so far as *Sara* is concerned (same weight, character, desires, reliable sense detectors) at the cost of a greater dissimilarity with respect to other facts obtaining at the time of the antecedent—facts about the state of the ice and the beliefs that Sara and others have about the state of the ice, etc. It seems plausible to suppose that the worlds which are *overall most similar* to our world at the time of the antecedent are worlds where the ice is in the same melted state it's actually in, and Sara is the same as she actually is except for the fact that she has somehow acquired the false belief that the ice is safe. If that's right, then the Fixed Law theory agrees (albeit for different reasons) with the counterfactual endorsed both by commonsense and the Naive theory: that if Sara had stepped on the ice, she would have fallen through.

But while the Fixed Law theory may give the correct truth-conditions for some agency counterfactuals, it's not clear that it works for all agency counterfactuals, either in terms of accounting for our knowledge of them or accounting for what we believe are the counterfactual facts. Here are two problems.

First, this theory allows us to consider particular facts obtaining at times earlier and later than the time of the antecedent only insofar as those facts may be deduced from the laws together with particular facts obtaining *at the time of the antecedent*. But it's not clear that this provides a sufficiently rich factual base to account for all the counterfactuals we think are true. For instance, in contemplating whether to climb a particular mountain, Sara might wonder whether she would be the first woman to do so. On the Fixed Law theory, there is an answer to this counterfactual question only if the present contains traces *either* of the fact that some woman previously climbed the mountain or of the fact that no woman has ever climbed the mountain. But it seems intelligible to suppose that the present provides us with no evidence either way. If so, then according to the Fixed Law theory the counterfactual "if Sara had climbed that mountain, she would have been the first woman to do so" is neither true nor false. But this is wrong for the same reasons that verificationist accounts of the past are wrong. Whether or not we can ever know it, it's either true or false that some woman has already climbed that mountain. And this historical fact suffices for the truth or falsity of "if Sara had climbed that mountain, she would have been the first woman to do so."

Second, if the laws must be held constant and determinism is true, then if anyone had done other than what she actually did, the causes of her choices would have been different, and the causes of those causes would have been different, and so on, all the way back to the Big Bang. And these would not be the only differences. For any world where our laws hold are worlds where these

different causes will have different effects, and these effects will in turn have different effects, and so on, all the way from the Big Bang back to the time of the antecedent of the counterfactual we are considering. Given this, the Fixed Law theory seems incapable of explaining our knowledge of even the most uncontroversial counterfactuals about the undone acts of deterministic agents. For instance, in knowing that if Dana had pushed the button, the light would have come on, we rely on our knowledge of many other counterfactuals, including counterfactuals about what would still be the case *at the time of the antecedent*. (The light bulb would still be working, the fuse would still not have blown, the cord would still not have come unplugged, and so on.) These counterfactuals *might* often be true, for every potentially causally relevant fact F. But there seems no guarantee, given the Fixed Law theory, that these counterfactuals will always be true. And if there is no guarantee, then the Fixed Law theory has not solved Goodman's problem; it has not given us a theory that makes it plausible that we have the counterfactual knowledge we seem to have.

These problems suggest a different kind of theory, one which is closer in some ways to the Naive theory. This theory—we'll call it the "Fixed Past" theory—tells us to evaluate agency counterfactuals by considering worlds where the past (prior to the person's choice) *is exactly the same* and the laws are, if need be, just different enough to allow the agent to choose differently. In other words, the laws are just different enough to allow the agent's choice to be an event that counterfactual theorists call "a divergence miracle".

The Fixed Past theory provides the right truth-conditions for the match counterfactual and for the various ice-stepping counterfactuals. But it also provides the right truth-conditions for counterfactuals, like the "first woman" counterfactual, that are true in virtue of historical facts. And since it holds the past constant, it's not subject to the "backtracking" worry that the Fixed Law theory is subject to. Finally, it provides a more plausible account of our knowledge of agency counterfactuals than does the Fixed Law theory. According to the Fixed Law theory, we are allowed to use our knowledge of the past (relative to the time of the antecedent) only insofar as this knowledge provides evidence about the state of the world at the time of the antecedent; according to the Fixed Past theory, we are allowed to rely on our knowledge of the past for the more straightforward reason that the past would be the same no matter what the agent does.

The Fixed Past theory looks pretty good. Is anything wrong with it? Well, if you were brought up on Goodman and the metalinguistic pre-possible worlds approach to counterfactuals, you may be suspicious of an account that seems to be playing "fast and loose" with the laws. Doesn't this lead to modal anarchy, you might wonder? If the laws at the closest antecedent-worlds are different from our own, who's to say what would happen at such worlds?

But this objection misunderstands how Lewis has taught us to think about the laws at the closest worlds.[30] Don't think of the closest worlds as worlds where one of our deterministic laws have been replaced by a different law, with

far-reaching consequences for different agents at different times. Think rather of worlds which share our past history until the occurrence of an event (eg. Sara's choosing to step on the ice) which, by the standards of *our laws*, is a "local miracle" and think of the laws at these "divergence miracle" worlds as being as much like our laws as is possible, given the occurrence of this event. At these worlds, one of our laws has been replaced by something that is, I think, best described as an "almost-law"; a generalization "weakened and complicated by a clause to permit the one exception".[31] To put it another way, the closest worlds, on this view, are worlds where events happen according to our laws at all places and times *except* in the small spatiotemporal region where the divergence miracle occurs.

Does this way of thinking about the laws at the closest worlds make any unacceptable assumptions either about the nature of laws or about possible worlds? I don't think so. No matter what your view of laws—whether you view them as true in virtue of the regularities that in fact obtain or whether you view them as true in virtue of relations of contingent necessitation—there are possible worlds where things happen in the way described above. And no matter what your view of possible worlds—whether you accept Lewis's robust realism about worlds or whether you think of worlds as abstract entities or useful fictions—what I described above is logically possible. It is possible that everything happens just as it actually does until the moment of Sara's choice, and that everything after her choice happens in accordance with our laws.

Given this, I see no reason for rejecting the Fixed Past theory of agency counterfactuals, and every reason for adopting it.

Let's review where we are. The success of the No Opportunity argument requires the truth of **DB**, and the truth of **DB** requires the truth of a theory of counterfactuals that tells us to always evaluate counterfactuals by holding the *laws* fixed. But while this is a natural assumption, it's equally natural to evaluate some counterfactuals—the ones I called "agency counterfactuals"—by holding the *past* fixed. If we combine these two natural assumptions, we end up with the Naive theory of agency counterfactuals. The Naive theory is a natural starting point, but it's naive insofar as it provides nontrivial truth-conditions for agency counterfactuals only if determinism is false (since if determinism is true, every world where someone does otherwise has either a different past or different laws). So we must reject the Naive theory and choose *either* a Fixed Law theory[32] or a theory that permits small divergence miracles as a trade-off for a greater match with respect to past and present particular fact. David Lewis has convinced many of us that the latter kind of theory is the right theory for the counterfactuals we entertain and assert in contexts where our primary concern is with answering questions about the causal upshots of the occurrence or non-occurrence of a particular event. I've been arguing that, so far as agency counterfactuals are concerned, the closest worlds are those where the *entire past* (prior to the agent's choice) is the same as it actually is. If I'm right about this, then **DB** is false and the No Opportunity argument fails.[33]

Ability, Opportunity, and Laws

We are now in a position to explain why our commonsense view of free will is, despite appearances, compatible with determinism. I have argued that our commonsense view embodies two elements; we assume that we have various mental skills or *abilities* (eg. to reason, deliberate, make choices, and so on) and we assume that when we act we ordinarily have the *opportunity* of choosing and doing otherwise, where this is understood so that it satisfies **FPA**. (If we had tried and succeeded in doing otherwise, the past prior to our choice would or at least might still have been just the same.) Finally, we assume that it's ordinarily true, at least on occasions when an agent deliberates, chooses, and acts, that she *could have done otherwise* in the following sense: she had both the ability to do something else and also the opportunity to do so.[34]

Given this view of free will, it's natural to suppose that if determinism is true, then we are *less* free than we would be if determinism were false (in the right kind of way). Determinism doesn't rule out the possession of any abilities, for we may have the ability to do something even when we are in circumstances where we lack the necessary conditions for exercising the ability (cf. the pianoless pianist). But if determinism is true, then it seems that we *can never do otherwise*; even if we have the relevant ability, we always lack the opportunity. For suppose that I have the ability to do some basic action X (eg. raise my hand) and suppose that on a particular occasion there is nothing that we would ordinarily count as an impediment to the exercise of my ability (I'm not in chains, unconscious, hypnotized, etc.). I consider doing X, but decide not to do it, and don't. Commonsense says that, barring unusual circumstances[35], I could have done X; in addition to the ability, I also had the opportunity. But if determinism is true, it seems that this belief must always be mistaken. For surely if I exercise (or try to exercise) any of my abilities, the laws of nature would still hold. So if I had done X, our deterministic laws would still have obtained and *the past would have been different*, indeed, *would have to have been different*. But if that's so, then appearances were misleading and I could not have done X; I had the ability, but lacked the opportunity.

This reasoning is natural and seductive, but I have argued that it rests on a mistake about counterfactuals. The mistake comes in supposing that we always evaluate counterfactuals about what we do (or try to do) by holding the laws fixed. Depending on the context and our interests, it *may* be appropriate to keep the laws fixed. For instance, when we try to answer the question: "Does S have the ability to do X?", it seems reasonable to suppose that we do so by considering the closest worlds where the laws are the same, where S is as similar to the way she actually is as is consistent with her lawfully trying to do X, and where the past is as different as it needs to be in order for it to be true that circumstances C obtain and S tries to do x. And when we seek to answer the question: "How might it come about that S would do X?", then it seems reasonable to suppose that we consider worlds where the laws are the same and S

is the same with respect to character, abilities, dispositions, and so on, and where the past is as different as it needs to be in order for it to be true that S lawfully, and in character, does X.[36] But in a situation where someone is trying to decide what to do by asking, of each considered act X, "What would be the causal consequences if I did X?", then, I argued, we consider worlds where the *past* prior to S's choosing to do X is exactly the same and the laws, if need be, are just different enough to accommodate S's choice. And in these contexts—agency contexts, as I called them—it's true that if S had done otherwise, the past prior to her choice would have been exactly the way it actually was. But then it's true, *even if determinism is true*, that we have, not just the ability to do something other than what we in fact do, but also the opportunity to do so. That is, we have free will in a sense that satisfies **FPA** *even if determinism is true*.

Is this enough to show that our commonsense view of free will is compatible with determinism? I think so. But my incompatibilist opponents are not convinced. They object for different reasons. Some are unconvinced by my claim about Fixed Past agency counterfactuals; they think that it's always true, in *every* relevant context of counterfactual utterance, that if a deterministic agent had done otherwise, the past would have been different. To these incompatibilists, I have nothing more to say except: "Show me a better theory of agency counterfactuals." Other incompatibilists are willing to accept the Fixed Past theory of agency counterfactuals, but they regard Commonsense Compatibilism (the claim that we have free will that satisfies **FPA** even if determinism is true) as an incredible thesis, even less plausible than standard varieties of compatibilism. It is to these incompatibilists that I address the remainder of my arguments in this section.

First, let's get clear about the source of the incredulity. In defending Commonsense Compatibilism, I am committed to the claim that if determinism is true, then there are occasions when it's true, of some agent S, and some action X, that:

> **Law-breaking Choice (LC):** S can do X and if she did X, her choice would be a law-breaking event; that is, the laws would be different and her choice would be the event in virtue of which this is true.

I grant that **LC**, considered in isolation, looks incredible, but I will argue that appearances are deceptive. I will argue that if you have agreed with me so far—that is, if you accept my claim about the abilities of deterministic agents, and you accept **FPA**, and you accept the Fixed Past theory of agency counterfactuals—then you should accept **LC**. **LC** is surprising, but it is neither incredible nor philosophically objectionable.

To see why, let's take a closer look at why it's natural to think that **LC** says something incredible.

We think that the laws constrain us, by setting limits on what we can do, but **LC** seems to deny this. **LC** seems to say that we can (if determinism is

true) choose in ways forbidden by the laws and act on these miraculous (law-breaking) choices. **LC** doesn't actually say that we can *do whatever* we like, despite the laws, but it seems arbitrary and ad hoc to say that we are able to make miraculous choices while denying that we are able to perform miraculous actions. But if we are able to *act* miraculously it seems that there are no limits at all to what we can do. This is not just incredible but clearly false. There are all sorts of things we cannot do—walk on water, run faster than the speed of light, defy gravity, etc. Since Commonsense Compatibilism has no grounds for denying that we can do these things, we must reject Commonsense Compatibilism.

There are two parts to this objection. The first part says that it's arbitrary to say that we are able to make law-breaking choices, while being unable to perform law-breaking actions. The second part says that Commonsense Compatibilism is committed to both claims and therefore denies that the laws place *any* restrictions on what we can do.

Let's look at the second part first. Commonsense Compatibilism says that we can do something only if we have the *ability* as well as the opportunity to do it, and Commonsense Compatibilism accepts the account of ability I proposed as common ground between compatibilists and incompatibilists. Someone has the ability to do X only if some worlds *with the same laws* where the person is in circumstances C and tries to do X are worlds where she succeeds in doing X. Given this account of what it is to have an ability, it follows that no one has the ability to do any act contrary to the laws. For there are no worlds with our laws at which anyone tries and succeeds in walking on water, running faster than the speed of light, or doing any other act which (either in itself, or together with circumstances C) entails the falsity of the laws. So the second part of the objection is answered. Commonsense Compatibilism does not deny that the laws place substantive constraints on what we can do; laws constrain us by setting limits on our abilities.

What about the first part of the objection, which says that it's arbitrary and ad hoc to say that agents have the ability to *choose* contracausally, while denying that they have the ability to *act* contracausally?

But Commonsense Compatibilism doesn't say that anyone has the *ability* to choose contracausally. Our doings include mental or psychological doings as well as doings of what are sometimes called basic actions (arm-raisings, foot-movings, etc.) and doings of nonbasic actions (riding a bicycle, playing the piano, etc.) Commonsense Compatibilism does not draw any arbitrary distinctions between our mental or psychological abilities and other abilities; the laws of nature constrain our mental abilities in exactly the same way they constrain our other abilities. We've got the ability to do a mental action just in case it's true that given the relevant circumstances C and given our laws, if we tried to do an act of the relevant mental type, we would probably succeed.

So both parts of this objection have been answered. Since we can do only what we have the ability to do, and since our abilities are constrained by the laws, it's false that Commonsense Compatibilism endorses the incredible claim

that the laws are irrelevant to questions of what we can do, including mental doings, including the doings which result in choices.

LC is misleading insofar as it seems to be attributing to S a certain kind of *ability*—the ability to make law-breaking choices. But **LC** must be read in the context of Commonsense Compatibilism, and, given this, what **LC** in fact asserts is a conjunction of two claims:

(**A**) S can, in the ability sense, do X

and

(**O**) If S now tried and succeeded in doing X, the past prior to her choice would be the same and so her choice would be the divergence miracle, that is, a law-breaking event.

The two conjuncts are logically independent of each other.

A is not sufficient for **O**. S may have the ability to do X without it also being true that if she now tried and succeeded in doing X, the past would be the same and her choice would be the divergence miracle. Unconscious Mary, tumbling Jack, and the pianoless pianist are examples.

And **O** is not sufficient for **A**. It may be true that if S now tried and succeeded in doing X, her choice would be the divergence miracle without it also being true that S has the ability to do X. For instance, suppose that S lacks the ability to pick the winning ticket out of the box, but that on a particular occasion nothing stands in the way of S getting lucky and picking out the winning ticket. If so, then it's true that *if* S had tried and succeeded in picking out the winning ticket, the past would have been the same until the occurrence of her choice which (if determinism is true) would have been a divergence miracle.

It shouldn't be surprising that **A** and **O** are independent of each other. **A** says that S has the ability to do X. **O** says, in part, that if S tried and succeeded in doing X, *the past prior to her choice would be exactly the same*; that is, **O** says that S has the opportunity to do X in the sense partly defined by **FPA**. Since we may have the opportunity to do something without having the ability to do it, we cannot draw any inference from the law-breaking counterfactual *also* asserted by **O** to the conclusion that S has any incredible abilities.

Libertarian Compatibilism

I've been calling my view "Commonsense Compatibilism", but I think it deserves a somewhat more provocative name. I hereby name it "Libertarian Compatibilism" because it is a compatibilist view which captures the intuitions behind our commonsense view, a view which is usually thought to be libertarian in the standard philosophical sense (that is, as entailing incompatibilism). We believe that we are agents, importantly different from the rest of

nature. We believe that we have the ability to transcend the forces that have made us what we are, that we are able somehow to "rise above" our desires and the chains of causation that bind the lower creatures. We believe that we enjoy, not unlimited freedom of the will, but a freedom that is absolute in this sense: barring very unusual circumstances, when we intentionally do something, we *could have chosen and done otherwise, given the actual past*. This commonsense view has traditionally been understood in the literature as the thesis that the laws which govern us are, at best, indeterministic, or more radically, that we aren't governed by laws at all but that we govern ourselves by a special brand of causation—agent-causation, where this is understood so that it reduces neither to event nor to fact causation.

What I have been arguing is that we can do justice to much in these intuitions without departing either from naturalism or from determinism. (I take no stand on whether determinism is in fact true or false, but follow the time-honored compatibilist tradition of saying it doesn't matter.) I've done this by arguing that what lies behind these intuitions is best understood counterfactually, in terms of **FPA**, and, more fundamentally, in terms of **ACA**. **FPA** says that we have free will only if it's true that if we did otherwise, *the past prior to our choice would or might still be the same*. **ACA** says that we have free will only if it's true that if we did otherwise, *the only differences would be our choice, action, and the causal consequences of our choice and action*. At first glance, it may seem that both **FPA** and **ACA** entail the incompatibility of free will and determinism; but I have argued that this is not the case. We have reasons independent of the free will/determinism debate for believing that the relevant counterfactuals are true.

Libertarian Compatibilism is an unorthodox form of compatibilism, but I think it's not only defensible but plausible. I think it captures what's intuitively right about compatibilism, on the one hand, and orthodox (that is, incompatibilist) libertarianism, on the other hand.

The following counterfactuals are consistent:

(C) If the past had been suitably different, S would have had different reasons and she would have chosen, tried, and succeeded in doing otherwise.

(L) If S had tried and succeeded in doing otherwise, the past prior to her choice would or at least might still have been exactly the same.

C is the claim traditionally stressed by compatibilists, who insist that a free agent is one whose actions casually and counterfactually depend on her reasons and whose reasons depend on facts about the past. **L** is the claim traditionally stressed by libertarians who insist that a person is free to do otherwise only if the past is counterfactually *independent* of her choice and action. But **C** and **L** are consistent, and, if I'm right about how we evaluate counterfactuals, then *both* are in fact sometimes true.[37]

Notes

1. The most sophisticated and rigorous arguments for incompatibilism are versions of the so-called "modal argument" articulated independently by Carl Ginet and Peter van Inwagen in a series of articles and in their books, Ginet's *On Action* (Cambridge, 1990) and van Inwagen's *An Essay on Free Will*, (Clarendon Press, Oxford, 1983). But these defenses of incompatibilism end up relying, at points that seem to me crucial, on undefended "intuitions" about the ways in which the past and laws are "fixed".
2. For a classic instance of this compatibilist strategy, see R.E. Hobart's "Free Will as Involving Determinism and as Inconceivable Without It", *Mind* 43 (1934), 1–27. For more recent examples, see Daniel Dennett's *Elbow Room: The Varieties of Free Will Worth Wanting*, Bradford Books, 1984, and Susan Wolf's *Freedom Within Reason*, Oxford University Press, 1990.
3. Van Inwagen articulates and defends three formal arguments for incompatibilism, but says that all three are versions of one basic argument. The one that's closest to the one I give here is the one he calls "The First Formal Argument". (*An Essay on Free Will*, ibid., pp. 68–78.)
4. Versions of this reply have been made by a number of philosophers, including John Fischer, "Incompatibilism", *Philosophical Studies* 43 (1983), 127–137 and David Lewis, "Are We Free to Break the Laws?", *Theoria* 47 (1981), 113–121.
5. Some people think that Harry Frankfurt has shown that being able to do otherwise is not a necessary condition of moral responsibility. (See his "Alternate Possibilities and Moral Responsibility", *Journal of Philosophy* 66 (1969), 829–839.) In my "Freedom, Foreknowledge, and the Principle of Alternate Possibilities" (*Canadian Journal of Philosophy*, forthcoming), I argue that so-called "Frankfurt stories" fail to establish this; if we have been so persuaded, it's because we have been taken in by a bad *argument*.
6. Although this is rough, it will do for our purposes. A satisfactory analysis of the ability to do X will have to be more carefully formulated. For recent discussion of the related project of trying to give a conditional analysis of dispositions, see C.B. Martin, "Dispositions and Conditionals", *Philosophical Quarterly* 44 (1994), 1–8, and David Lewis, "Finkish Dispositions", *Philosophical Quarterly* 47 (1997), 143–158.
7. On the contrary, a venerable compatibilist strategy has been to argue that the possession of the relevant abilities requires the truth of determinism. This strategy fails because it assumes that there is no causation in the absence of determinism. It's now generally accepted that there may be causation—understood either in terms of subsumption under probabilistic laws or in terms of probabilistic counterfactual dependence—even if determinism is false. Given this, it's false that an undetermined event is for that reason a random event, not in anyone's causal control. And it's false that the falsity of determinism entails that there are no abilities.
8. Why "at least sometimes" instead of 'always'? Because it would be implausible to understand the incompatibilist thesis as the claim that free will requires that we are always free to do otherwise. It seems implausible to suppose that we cease to have free will each time we fall asleep even though our state of unconsciousness prevents us from exercising any of our unexercised abilities. Given this, we have to understand the incompatibilist as *either* saying that we have free will only if we at least sometimes have both ability and opportunity to do otherwise or as saying that

we have free will only if it's ordinarily true, on *each occasion of choice or action*, that we have both ability and opportunity to do otherwise.

9. For "choice", you may substitute "decision", "intention", "volition" or whatever you think the causal antecedent (or first part) of action is. I remain neutral in this paper about different theories of action; my arguments can be reformulated so they apply regardless of your view about what actions are.
10. Note that **ACA** is weak in several ways. It states only a necessary condition for having free will, and it allows for the possibility that someone has free will even if there are relatively few occasions on which she has opportunity as well as ability to do otherwise. I've formulated **ACA** so weakly for two reasons: First, to avoid the unattractive consequence that we cease to have free will whenever we fall asleep. (See note 8.) Second, because I want to capture the common core of a wide range of different views about what it takes for someone to have free will; these differences are, I think, best understood as different views about what *abilities* are necessary for free will.
11. Tell any story where you believe that all of the following are true: Someone cannot do X; if she chose (tried, etc.) to do X, she would succeed; she cannot do X because she suffers from some state (a pathological aversion, phobia, extreme panic, hypnosis, etc.) which renders her unable to choose, decide, intend or in any way try to bring it about that she does X. The moral I draw applies to any story that meets these conditions.
12. Recall our discussion of the Incredible Ability argument, in which we distinguished the innocuous "backtracking" C1 from the incredible causal "backtracking" C2 and argued that the compatibilist is committed only to C1.
13. I don't mean to imply that it cannot be resisted. It can be resisted by providing a compatibilist account of "is able to do x". A promising beginning is the idea that someone is able to do X provided that the following are true: i) she has the ability to do X; ii) she has the relevant mental abilities and capacities—e.g. the ability to deliberate concerning the reasons for and against doing X, the ability to form a judgment based on her consideration of the reasons for and against doing X, and the ability to act according to her judgment; iii) none of the relevant abilities or capacities are impaired or malfunctioning (e.g.. due to drugs, hypnosis, extreme panic, etc.) and iv) there is no external impediment to her doing x.
14. See Fischer's *The Metaphysics of Free Will*, Blackwell, 1994, p. 178. Harry Frankfurt's famous argument for the thesis that we may be morally responsible even if determinism renders us unable to do otherwise was based on a thought experiment involving an agent who does something for his own reasons and who is intuitively responsible for what he does despite the existence of a powerful being, in the background, who would prevent him from acting in any other way. ("Alternate Possibilities and Moral Responsibility", *ibid.)*
15. P.J. Downing first directed our attention to the fallacious "backtracking" argument rejected by Sara in "Subjunctive Conditionals, Time Order, and Causation", *Proceedings of the Aristotelian Society* 59 (1958–59), 125–140. The fallacious argument is: "If Sara stepped on the ice, then it would not have melted this morning; if the ice had not melted this morning and Sara stepped on it, she would not fall through; therefore, if Sara stepped on the ice, she would not fall through."
16. If Bizet and Verdi had been compatriots, would they both have been Italian or would they both have been French? Is there a fact of the matter or does it all depend on what similarity respects matter most to the person considering the counterfactual?

If we focus only on counterfactuals of the Bizet/Verdi kind, we may be tempted to conclude that counterfactuals *never* have objective truth-conditions. But this conclusion would be, I think, premature. There are many different kinds of counterfactuals; it would not be surprising if some have, while others lack, objective truth-conditions.

17. "The Problem of Counterfactual Conditionals", *Journal of Philosophy* 44 (1947), 113–128. Reprinted as Chapter 1 of *Fact, Fiction, and Forecast*, Cambridge, Mass., 1984.
18. This way of putting it is mine, not Goodman's. But from his examples it's clear that he was thinking of counterfactuals which are *causal* (the fact or event referred to by the antecedent would be or would have been a cause of the fact or event referred to by the consequent) and *singular* as opposed to general insofar as they are uttered on some particular occasion with particular background conditions. (Eg. "if this match had been scratched, it would have lit"; "if that radiator had frozen, it would have broken:; "if that piece of butter had been heated to 150 F, it would have melted".)
19. I will use "choice" to refer to whatever mental event is either the cause or the first part of an agent's intentional act. You may substitute "intention", "decision", "volition", or whatever you think the relevant mental event is.
20. A "divergence miracle" is an event that is unlawful by the standards of *our* laws in the following sense: the conjunction of some earlier facts together with the fact that the event occurred entails the falsity of our laws. There are no events that are unlawful by the standards of the world at which they occur, so the worlds where the divergence miracle occurs are worlds where the laws are slightly different from our laws.
21. See, for instance, "Counterfactual Dependence and Time's Arrow", *Nous* 13 (1979), 455–476, reprinted, with Postscripts, in his *Philosophical Papers* Vol II, Oxford, 1986, pp.32–66. Lewis doesn't claim to be giving an account of singular causal counterfactuals; he makes the more ambitious claim that this is how we evaluate counterfactuals given what he calls "the standard resolution" for counterfactual vagueness. (*Philosophical Papers*, pp.33–34) As I read Lewis, the "standard resolution" includes, but is not necessarily limited to, those counterfactuals we entertain in contexts where our primary interest in in figuring out what the *causal upshots* of some particular event or action would be. Before giving his theory, Lewis acknowledges that there are some special contexts where we evaluate counterfactuals differently, and he gives as an example a counterfactual of the "if Sara had stepped on the ice at noon, then (given her cautious character), the ice would not have melted this morning" variety. He offers no theory for these "non-standard" counterfactuals.

    I'm not sure whether Lewis is right in drawing the standard/nonstandard distinction in the way that he does, but I think that he is right about that subset of his "standard" counterfactuals in which our main concern is in figuring out the causal consequences of a particular event or action.
22. In saying this, I take myself to be saying something that Lewis would not (or should not) deny. That is, I think that if we apply Lewis's theory of "standard resolution" counterfactuals to the special case of agency counterfactuals, then the closest worlds where the agent intentionally does otherwise will always turn out to be worlds where the agent's choice (or intention, decision, etc.) is the divergence miracle. It should be noted, however, that Lewis does not make this claim in his discussion of the free will/determinism problem in "Are We Free to Break the Laws?", *ibid*. What he says

there is that *if* determinism is true, then if he had done otherwise, for instance, if he had raised his hand, then "the course of events would have diverged from the actual course of events a little while before I raised my hand, and at the point of divergence, there would have been a law-breaking event—a divergence miracle". This is consistent with the divergence miracle being the agent's choice, but it seems to leave it open that the miracle is some other, perhaps slightly earlier event.

23. Possible worlds semantics for counterfactuals was developed independently by Robert Stalnaker ("A Theory of Conditionals", *Studies in Logical Theory, American Philosophical Quarterly*, Monograph 2, Blackwell, 1968, pp. 98–112) and David Lewis (*Counterfactuals*, Harvard University Press, 1973) The use of possible worlds semantics as a tool for evaluating counterfactuals is neutral on different solutions to Goodman's problem, neutral on the question of whether counterfactuals have objective truth-conditions, and neutral on different conceptions of possible worlds.
24. This formula should be understood so that "closest" means "more close than any other" and it should not be assumed that closeness is a matter of degree. It's generally agreed that the closest worlds are similar to the actual world, but not everyone attempts to define closeness in terms of similarity.
25. I put it this way, in terms of events, to highlight that the counterfactuals under considerations are all counterfactuals where the antecedent refers to something that has causes and effects. Most philosophers think that causal relata are events, so that's how I've put it. If you think that causal relata are facts, feel free to substitute "if fact F", keeping in mind that F is the kind of fact that can be a causal relata.
26. See, for instance, G.H. Von Wright, "Causality and Causal Explanation", in *Explanation and Understanding*, Cornell University Press, 1971.
27. See for instance, Michael Dummett, "Bringing about the Past", *The Philosophical Review* 73 (1964), 338–359. See also David Lewis, "The Paradoxes of Time Travel", *American Philosophical Quarterly* 12 (1976), 145–152, and Kadri Vihvelin, "What Time Travelers Cannot Do", *Philosophical Studies* 81 (1996), 315–330.
28. There are different versions of what I'm calling "the Fixed Law" theory. What they have in common is that they try to solve Goodman's problem within the confines of a theory that says that the closest worlds all have the same laws as our world. The version I give below is, more or less, the one articulated by Jonathan Bennett in "Counterfactuals and Temporal Direction", *The Philosophical Review* 93 (1984), 57–91. (Bennett no longer endorses this theory; I describe it because it's the best version of a Fixed Law theory that I know.) See also Paul Horwich, chapter 10 of *Asymmetries in Time*, Bradford Books, MIT Press, 1989.
29. "Counterfactuals and Temporal Direction", p. 73, ibid.
30. See his *Counterfactuals,* ibid, p. 75 and "Counterfactual Dependence and Time's Arrow", in *Philosophical Papers* II, pp. 38–56, ibid.
31. *Counterfactuals,* p.75, ibid.
32. I criticised just one version of a Fixed Law theory—the version defended (but no longer endorsed—see note 28) by Jonathan Bennett. However, I think that similar objections will prove fatal to any attempt to provide a Fixed Law theory for *agency* counterfactuals. In saying this, I don't mean to imply that it's *never* appropriate to hold the laws fixed when we evaluate counterfactuals.
33. In making these arguments, I've appealed only to considerations about the evaluation of counterfactuals. I have not assumed that a deterministic agent has free will in the sense at issue in the free will/determinism debate. An incompatibilist might agree with everything I've said so far about counterfactuals, and still deny that any

deterministic agent can ever do other than what she does. Note, however, that such an incompatibilist needs to defend her position by way of something other than the No Opportunity argument.

34. The qualification "ordinarily" is necessary because there are unusual cases, involving "backup" interveners of the sort described in some Frankfurt stories (see notes 5 and 14) in which someone deliberates, chooses, and acts intentionally but could not have *successfully* acted in any other way.
35. For instance, circumstances of the sort described in some Frankfurt stories. See notes 5, 14, and 34.
36. This is the counterfactual question that we are guided by when we assert or entertain counterfactuals like "if Sara had stepped on the ice, she would have checked first to make sure it's safe" and "we can be sure that if Sara ever stepped on ice, she would not fall through".
37. I am grateful to Mark Balaguer, Jonathan Bennett, Mark Bernstein, Robert Bright, Curtis Brown, Randolph Clarke, John Fischer, Pieranna Garavaso, Carl Ginet, Ishtiyaque Haji, Mark Heller, Hud Hudson, Robert Kane, Tomis Kapitan, Barry Loewer, Michael Otsuka, Howard Sobel, Terrance Tomkow, Gideon Yaffe, and Michael Zimmerman for helpful comments on earlier versions of this paper.