CHAPTER 11

WHO'S AFRAID OF DETERMINISM? RETHINKING CAUSES AND POSSIBILITIES

CHRISTOPHER TAYLOR
DANIEL DENNETT

INCOMPATIBILISM, the view that free will and determinism are incompatible, subsists on two widely accepted but deeply confused theses concerning possibility and causation: (1) In a deterministic universe, one can never truthfully utter the sentence "I could have done otherwise," and (2) In such universes, one can never really take credit for having caused an event, since in fact all events have been predetermined by conditions during the universe's birth. Throughout the free will literature, one finds variations on these two themes, often intermixed in various ways. When Robert Nozick[1] describes our longing for "originative value," he apparently has thesis (2) in mind, and thesis (1) may underlie his assertion that "we want it to be true that in that very same situation we could have done (significantly) otherwise." John Austin, in a famous footnote, flirts with thesis (1):

> Consider the case where I miss a very short putt and kick myself because I could have holed it. It is not that I should have holed it if I had tried: I did try, and missed. It is not that I should have holed it if conditions had been different: that might of course be so, but I am talking about conditions as they

> precisely were, and asserting that I could have holed it. There is the rub. Nor does "I can hole it this time" mean that I shall hole it this time if I try or if anything else; for I may try and miss, and yet not be convinced that I could not have done it; indeed, further experiments may confirm my belief that I could have done it that time, although I did not.[2]

(In later sections we discuss at length the ways in which this particular quote can lead readers astray.) Meanwhile, Robert Kane, in *The Significance of Free Will*, eloquently proclaims the importance of our presumed ability truly to cause events, the ability that thesis (2) addresses:

> Why do we want free will? We want it because we want ultimate responsibility. And why do we want that? For the reason that children and adults take delight in their accomplishment from the earliest moments of their awakening as persons, whether these accomplishments are making a fist or walking upright or composing a symphony.[3]

Elsewhere in the free will debate, one often finds authors advancing definitions that confirm the relevance of possibilities and causes. Kane describes free will itself, for instance, as "the power of agents to be the ultimate creators . . . and sustainers of their own ends and purposes."[4] The key words here are *power* and *creator*. Intuition suggests that the term *power* is intertwined with *possibility* roughly as follows: agent A has the power to do X if and only if it is possible that A does (will do) X. And certainly, to be a *creator*, one has to be the *cause* of changes in the world; one has to "make a difference" in how the world runs. Kane provides some other significant concepts:

> Alternative Possibilities (AP): The agent has *alternative possibilities* . . . with respect to A at t [iff] at t the agent can (has the *power* or *ability* to) do A and can do otherwise.[5]

> Ultimate Responsibility (UR): An agent is *ultimately responsible* for some (event or state) E's occurring only if (R) . . . something the agent voluntarily . . . did or omitted and for which the agent could have voluntarily done otherwise . . . causally contributed to E . . . , and (U) for every X and Y . . . [I]f the agent is personally responsible for X, and if Y is an *arche* (or sufficient ground or cause or explanation) for X, then the agent must also be personally responsible for Y.[6]

Carl Ginet in a similar vein proposes:

> Two or more alternatives are *open to me* at a given moment if which of them I do next is entirely up to my choice at that moment: Nothing that exists up to that moment stands in the way of my doing next any one of the alternatives.[7]

Whether or not these definitions are entirely dependable, they are emblematic of the central role of the concepts of causation and possibility in our understanding of free will.

In short, the acceptance of theses (1) and (2) lies at the heart of incompatibilism. Incompatibilists dread determinism because they suspect that a deterministic universe would lack the sorts of open possibilities that we cherish and deprive us of the ability to cause changes to the world in a meaningful way. Accordingly, they find heartening the discovery of indeterminacy in modern quantum mechanics, and they hope to discover indeterministic quantum events at the root of each free agent's decision-making ability. Kane ingeniously attempts a naturalistic, scientifically respectable account of indeterministic free will, and yet the arcane processes he describes are strangely dissatisfying as a new foundation for human freedom and dignity. Not only do they seem oddly "outside of our control," but they are so subtle that, very likely, scientists will be unable to confirm their relevance to our mental life for the foreseeable future.

To avoid the sort of impasse that Kane and other incompatibilists have apparently reached, we propose to reexamine the foundations of possibilities and causes, to understand why theses (1) and (2) look so compelling. We will discover that the desires incompatibilists describe, to have powers and to effect changes, can be satisfied without any recondite appeals to quantum indeterminacy. The suspicions to the contrary lose their force once we begin to untangle, with the aid of a little formalism, the complexities of the underlying concepts.

1. Possible Worlds

While a complete account of possible worlds would require many extra pages, the following paragraphs outline an approach, compatible with modern scientific methods, that avoids various modal pitfalls identified by Quine (such as talk of *propositions, analyticity, essences*, and so forth).[8] Ideally, science strives for a description of the universe that is as thorough and comprehensive as possible, composed in an orderly mathematical idiom. A simple example of such ideal state-descriptions are the "Democritean" universes introduced by Quine.[9] A Democritean universe is completely specified using a function f that assigns to each quadruple (x,y,z,t) a value of either 0 or 1. If $f(x,y,z,t) = 1$, then at time t matter occupies location (x,y,z); otherwise point (x,y,z) is devoid of matter at t. Needless to say, modern physics has long since supplanted the tidy Democritean conception of reality, but even today the basic project of describing the world with (monstrously complex) functions remains intact. So despite its scientific shortcomings, the following definition provides a useful starting point as we struggle to discipline unruly pretheoretical intuitions:

> A *possible world* is simply any function of the form described above (in mathematical notation, any function of the form f: $\mathbb{R}^4 \rightarrow \{0,1\}$).[10]

The set of all possible worlds we will denote by Ω a particularly noteworthy subset of Ω is Φ, which contains just the *physically* or *nomologically possible* worlds, in which no physical laws are violated.[11]

Given a possible world f, we of course have many ways to describe and make assertions about it. Often it will be natural to postulate *entities* within f: connected hypersolids in $\mathbb{R}^4$ that yield coherent life-histories for objects like stars, planets, living creatures, and everyday paraphernalia. One will also want to set up a system of *informal predicates* that apply to these entities, such as "has a length of 1 meter," "is red," "is human."[12] We may then form sentences like

> $\exists x$ (x is human)

and determine whether they apply in various different possible worlds (while recognizing that often enough one will encounter borderline worlds where incontestible verdicts prove elusive).

Worthy of special note are *identification predicates* of the form "is Socrates." "Is Socrates," we shall suppose, applies to any entity in any possible world that shares so many features with the well-known denizen of the actual world that we are willing to consider it "the same person." In the actual world, of course, "is Socrates" applies to exactly one entity; in others, there may reside no such being, or one, or conceivably several to whom the predicate applies equally well. Like other informal predicates, identification predicates suffer from vagueness and subjectivity, but they do not cause unusual problems.

With this machinery in place we can now explicate such sentences as:

> Necessarily, Socrates is mortal. (1)

We would propose the translation:

> In every (physically?) possible world f, the sentence "$\forall x$ (x is Socrates $\Rightarrow$ x is mortal)" obtains. (2)

Here "is Socrates" and "is mortal" are informal predicates of the sort just introduced. Paraphrase (2) strikes us as both plausible and free of the logical confusions Quine decries. Of course, deciding whether (2) is true does present considerable challenges, stemming largely from the unavoidable blurriness of the predicates. Moreover, we are not specifying the set of possible worlds over which one should allow f to range; perhaps some readers will advocate set Ω (all worlds), others Φ (the physically possible worlds), and yet others a still more restricted set X. Logic

alone cannot resolve this issue, but logical language does help us to pinpoint such questions and recognize the sorts of vagueness we face. However we choose, we can employ the notation

$$\Box_X \phi$$

to indicate that sentence φ obtains for every world in set X.

As the dual of necessity, possibility yields to a similar analysis. Hence

Possibly, Socrates might have had red hair. (3)

means

There exists (within some set X) a possible world f in which the sentence "$\exists x$ (x is Socrates $\wedge$ x has red hair)" obtains. (4)

Analogous to the notation "$\Box_x \varphi$" we introduce

$$\Diamond_X \varphi,$$

meaning that φ holds for some world within X. The familiar sentence:

Austin could have holed the putt (5)

now becomes

$$\Diamond_X \exists x \ (x \text{ is Austin} \wedge x \text{ holes the putt}). \quad (6)$$

Notice that in this case we need to restrict X to a narrow range of worlds, all quite similar to actuality, if we are to do justice to Austin's meaning. For suppose that Austin is an utterly incompetent golfer, and that impartial observers are inclined to deny (5). If we let X range too widely, we may include worlds in which Austin, thanks to years of expensive lessons, winds up a championship player who holes the putt easily, thus validating (6) but distorting the presumed sense of (5). At the same time, as we shall see, there is no good reason to make X so small that only worlds *identical* to reality in the moments before the putt are included.

2. COUNTERFACTUALS

Using possible worlds, one can also profitably interpret sentences of the form

If you had tripped Arthur, he would have fallen, (7)

as David Lewis has shown.[13] Roughly, (7) obtains if and only if in every world approximately similar to our own where the antecedent holds, so does the consequent. In other words,

$$\Box_X \varphi \Rightarrow \psi, \tag{8}$$

where φ stands for "you tripped Arthur," ψ stands for "Arthur fell," and X is a set of worlds similar to our own. As an alternative notation, let us also write,

$$\varphi \,\Box\!\!\rightarrow_X \psi. \tag{9}$$

Choosing an optimal value for X in (8) and (9) is not always easy, but we suggest the following loose guidelines:

In sentences like (8) and (9), X ought to

- contain worlds in which φ holds, $\sim\varphi$ holds, ψ holds, and $\sim\psi$ holds
- contain worlds otherwise very similar to the actual world (insofar as the preceding clause permits). (G)

So when analyzing (7), choose X to contain worlds in which you trip Arthur, worlds where you refrain from tripping him, worlds where he falls, and worlds where he remains upright. In the case of (10):

If the sun hadn't risen this morning, I would have overslept, (10)

X will look quite different, since it includes strange worlds in which the sun fails to rise.

In *Counterfactuals*, Lewis cleverly devises a single connective $\Box\!\!\rightarrow$ appropriate for all φ and ψ, but in this chapter we settle for a family of connectives of type $\Box\!\!\rightarrow_x$. Doing so, we believe, forestalls various technical complications and accords equally well with intuition. Notice that for Lewis transitivity fails and, worse, so does the equivalence

$$\varphi \,\Box\!\!\rightarrow \psi \equiv \sim\psi \,\Box\!\!\rightarrow \sim\varphi.$$

With each operator $\Box\!\!\rightarrow_X$, on the other hand, transitivity and contraposition succeed, provided we hold X fixed. Of course, X can vary, as observed in the previous paragraph, so that the two sentences:

If Bill had tripped him, he would have fallen (11)
If he had fallen, he would have broken his glasses (12)

need not imply

If Bill had tripped him, he would have broken his glasses. (13)

However, we can confidently assume that

If Bill had tripped him, he would have fallen (14)

implies

If he had not fallen, Bill would not have tripped him, (15)

since guidelines (G) yield the same set in each case.

3. CAUSATION

Fundamental as it appears, the language of causation has stirred up interminable debate and has (perhaps for that reason) been avoided by scientists. Many philosophers apparently hope some day to unearth the one "true" account of causation, but given the informal, vague, often self-contradictory nature of the term, we think a more realistic goal is simply to develop a formal analogue (or analogues) that helps us think more clearly about the world. Our preexisting hunches about causation will provide some guidance, but we should mistrust any informal arguments that masquerade as "proofs" validating or debunking particular causal doctrines.[14]

When we make an assertion like

Bill's tripping Arthur caused him to fall, (16)

a number of factors appear to be supporting the claim. In an approximate order of importance, we list the following:

- Causal necessity. At least since Hume, philosophers have suspected that counterfactuals play some role in our causal thinking, and this factor and the next fall within the same tradition. Our assent to sentence (16) depends on our conviction that in any world roughly similar to our own in which Arthur falls, Bill must have tripped him up. Using the notation of the previous section, we have $\psi \,\Box\!\!\rightarrow_X \varphi$, where φ stands for "Bill tripped Arthur," ψ represents "Arthur fell," and X is the set of worlds similar to our own in which (1) Bill trips Arthur, (2) Bill doesn't trip him, (3) Arthur falls, or (4) he doesn't fall. As observed above, the sentence $\sim\varphi \,\Box\!\!\rightarrow_X \sim\psi$ has the same logical force; in other words, had Bill not tripped Arthur, he would not have fallen.
- Causal sufficiency. It may well be that whenever we affirm (16), we do so partly because we believe that (using the same notation as before) $\varphi \,\Box\!\!\rightarrow_X \psi$. In other words, we believe that Arthur's fall was an *inevitable outcome* of Bill's tripping: in any world where Bill places the obstruction in his path, Arthur goes toppling. (Or equivalently, if Arthur had *not* fallen, then Bill must in that case have refrained.) This second condition is logically entirely distinct from the first, and yet the two seem to get badly muddled in everyday thinking. Indeed, as we shall see, incompatibilist confusion often originates precisely here. Below we will discuss at greater length the relations between these two conditions.
- Truth of φ and ψ in the actual world. Although a relatively trivial requirement, it should be mentioned if only for completeness.
- Independence. We expect the two sentences φ and ψ to be logically independent: there must exist worlds, however remote from reality, in which φ obtains but not ψ, and vice versa. Hence "Mary's singing and dancing caused her to dance and sing" has a decidedly odd ring. This condition also helps rule out "$1 + 1 = 2$ causes $2 + 2 = 4$."
- Temporal priority. A reliable way to distinguish causes from effects is to note that causes occur earlier.[15]
- Miscellaneous further criteria. Although less critical than the preceding points, a number of other conditions may increase our confidence when we make causal judgements. For instance, in textbook examples of causation, φ often describes the actions of an agent, and ψ represents a change in the state of a passive object (as in "Mary causes the house to burn down"). Further, we often expect the two participants to come into physical contact during their transaction.[16]

In order to understand these conditions better, let us try them out on a few test cases (some of which derive from Lewis).[17] First consider the sharpshooter aiming at a distant victim. Scrutiny of the sharpshooter's past record shows that

the probability of a successful hit in this case is 0.1; if it makes any difference, we might imagine that irreducibly random quantum events in the sharpshooter's brain help determine the outcome. Let us suppose that in the current case the bullet actually hits and kills the victim. We unhesitatingly agree then that the sharpshooter's actions caused the victim's death, *despite their causal insufficiency.* Accordingly, it appears that in such cases, people rank necessity above sufficiency when making judgments about causes.

Still, sufficiency does retain some relevance. Suppose that the king and the mayor both have an interest in the fate of some young dissident; as it happens, both issue orders to exile him, so exiled he is. This is a classic case of *overdetermination.* Let φ_1 stand for "the king issues an exile order," φ_2 stand for "the mayor issues an exile order," and ψ, "the dissident goes into exile." In the current scenario, neither φ_1 nor φ_2 alone is necessary for ψ: for instance, had the king failed to issue any order, the dissident would still have been exiled thanks to the mayor, and vice versa. In fact $\varphi_1 \wedge \varphi_2$ satisfies the necessity requirement, but we are (perhaps unreasonably) reluctant to posit a disjunction as a cause.[18] Instead, sufficiency comes to the rescue and permits a choice between the two. After all, φ_2 fails this test: it is easy to imagine a universe where the mayor issues his decree yet the dissident gets off (just change the king's order into a pardon). The king's order, on the other hand, is truly *effective*; whatever small changes we make to the universe (including changes in the mayor's orders), the dissident's exile follows from the king's command. Accordingly we may dub φ_1 the "real cause" (if we feel the need to satisfy that yearning).[19]

Consider next the tale of Billy and Susie. Both children are throwing rocks at a glass bottle, and as it happens Susie's rock, traveling slightly faster, reaches the bottle first and shatters it. Billy's rock arrives a moment later at exactly the spot where the bottle used to stand, but of course encounters nothing but flying shards. When choosing between φ_1 ("Susie throws rock S") and φ_2 ("Billy throws rock B"), we vote for φ_1 as the cause of ψ ("The bottle shatters"), despite the fact that neither sentence is necessary (had Susie not thrown her rock, the bottle would still have shattered thanks to Billy, and vice versa) and both are sufficient (Billy's throw suffices to produce a broken bottle, whatever his playmate does, and likewise with Susie's). Why? The general notion of temporal priority (introduced above in connection with distinguishing cause from effect) strikes us as one critical consideration. As with priority disputes in science, art, and sports, we seem to put a premium on being the *first* with an innovation, and since rock S arrived in the vicinity of the bottle earlier than did rock B, we give credit to Susie. Further, it is clear that, although the bottle would still have shattered without Susie's throw, the shattering event would have been significantly different, occurring at a later time with a different rock sending fragments off in different directions. We can choose set X to reflect this fact (in keeping with guidelines (G)):

let it contain worlds in which either (1) the bottle doesn't shatter at all, or (2) it shatters in a way very similar to the way it shatters in reality. Then for every world in X,

$$\psi \Rightarrow \varphi_1$$

obtains; wherever in X the bottle shatters, we find Susie throwing her rock first. On the other hand,

$$\psi \Rightarrow \varphi_2$$

may well fail in X; X can certainly contain worlds where the bottle shatters but Billy refrains. In short, φ_1 is "more necessary" than φ_2, provided that we choose X right. The vagueness of X, though sometimes irksome, can also break deadlocks.

Not that deadlocks must always be breakable. We ought to look with equanimity on the prospect that sometimes circumstances will fail to pinpoint a single "real cause" of an event, no matter how hard we seek. A case in point is the classic law school riddle:

> Everybody in the French Foreign Legion outpost hates Fred, and wants him dead. During the night before Fred's trek across the desert, Tom poisons the water in his canteen. Then, Dick, not knowing of Tom's intervention, pours out the (poisoned) water and replaces it with sand. Finally, Harry comes along and pokes holes in the canteen, so that the "water" will slowly run out. Later, Fred awakens and sets out on his trek, provisioned with his canteen. Too late he finds his canteen is nearly empty, but besides, what remains is sand, not water, not even poisoned water. Fred dies of thirst. Who caused his death?[20]

4. Determinism and Possibility (Thesis 1)

Now that we have some formal machinery in place, we can reconsider the spuriously "obvious" fear that determinism reduces our possibilities. We can see why the claim *seems* to have merit: let φ be the sentence "Austin holes the putt," let X be the set of physically possible worlds that are *identical* to the actual world at some time t_0 prior to the putt, and assume both that Austin misses and that determinism holds. Then in fact φ does not hold for any world in X ($\sim \Diamond_X \varphi$), because X contains only one world: the actual one. Of course, this method of choosing X (call it the *narrow method*) is only one among many. We should note

that the moment we admit into X worlds that differ in a few imperceptibly microscopic ways from actuality at t_0, we may well find that $\Diamond_X \varphi$, even when determinism obtains. (This is, after all, what recent work on chaos has shown: many phenomena of interest to us can change radically if one minutely alters the initial conditions.) So the question is: when people contend that events are possible, are they really thinking in terms of the narrow method?

Notice that Austin evidently endorses the narrow method of choosing X when he states that he is "talking about conditions as they precisely were" whenever he asserts he could have holed the putt. Yet in the next sentence he seemingly rescinds this endorsement, observing that "further experiments may confirm my belief that I could have done it that time, although I did not." What "further experiments" might indeed confirm Austin's belief that he could have done it? Experiments on the putting green? Would his belief be shored up by his setting up and sinking near-duplicates of that short putt ten times in a row? If so, then he is not as interested as he claims he is in conditions as they precisely were. He is content to consider "Austin holes the putt" possible if, in situations very similar to the actual occasion in question, he holes the putt.[21]

We contend, then, that Austin equivocates when he discusses possibilities, and that in truth the narrow method of choosing X does not have the significance he imagines. From this it follows that the truth or falsity of determinism should not affect our belief that certain unrealized events were nevertheless "possible," in an important everyday sense of the word. We can bolster this last claim by paying a visit to a narrow domain in which we know with certainty that determinism reigns: the realm of chess-playing computer programs.

Computers are marvels of determinism. Even their so-called random number generators only execute pseudo-random functions, which produce *exactly* the same sequence of "random" digits each time the computer reboots. That means that computer programs that avail themselves of randomness at various "choice" points will nevertheless spin out exactly the same sequence of states if run over and over again from a cold start.[22] Suppose, for instance, you install two different chess-playing programs on your computer and yoke them together with a little supervisory program that pits them against each other, game after game, in a potentially endless series. Will they play the same game, over and over, until you turn off the computer? Perhaps; but if either chess program consults the random number generator during its calculations (if, for instance, it periodically "flips a coin" to escape from Buridan's ass difficulties in the course of its heuristic search), then in the following game the state of the random number generator will have changed. Accordingly, different alternatives will be "chosen" and a variant game will blossom, resulting in a series in which the games, like snowflakes, are no two alike.[23] Nevertheless, if you turned off the computer and then restarted it, running the same program, exactly the same variegated series of games would spin out.

This gives us a toy model of a deterministic Democritean universe, in which zillions of bits are flipped in sequence, governed by a fixed physics. Rewinding and replaying the tape of life is really possible in such a toy world. Suppose we create such a chess universe involving two programs, A and B, and study the results of a lengthy run. We will find lots of highly reliable patterns. Suppose we find that A (almost) always beats B. That is a pattern that we will want to explain, and saying, "Since the program is deterministic, A was *caused* always to beat B" would fail to address that curiosity. We will want to know what about the structure, methods, and dispositions of A accounts for its superiority at chess. A has a competence or power that B lacks, and we need to isolate this interesting factor.[24] When we set about exploring the issue, availing ourselves of the high-level perspective from which the visible "macroscopic" objects include representations of chess pieces and board positions, evaluations of possible moves, decisions about courses to pursue, and so forth, we will uncover a host of further patterns: some of them endemic to chess wherever it is played (for example, the near certainty of B's loss in any game where B falls a rook behind) and some of them peculiar to A and B as particular chess players (for example, B's penchant for getting its queen out early).[25] We will find the standard patterns of chess strategy, such as the fact that when B's time is running out, B searches less deeply through the game tree than it does when in the same position it has more time remaining. In short, we will find a cornucopia of *explanatory* regularities, some exceptionless (in our voluminous run) and others statistical.

These macroscopic patterns are salient moments in the unfolding of a deterministic pageant that, looked at from the perspective of microcausation, is to a large extent all the same. What from one vantage point appear to us to be two chess programs in suspenseful combat, can be seen through the "microscope" (as we watch instructions and data streaming through the CPU) to be a single deterministic automaton unfolding in the only way it can, its jumps already predictable by examining the precise state of the pseudo-random number generator. There are no "real" forks or branches in its future; all the "choices" made by A and B are already determined. Nothing, it seems, is really *possible* in this world other than what actually happens. Suppose, for instance, that an ominous mating-net looms over B at time t but collapses when A runs out of time and terminates its search for the key move one pulse too soon; that mating net *was never going to happen.*[26] (This is something we could prove, if we doubted it, by running the same tournament another day. At exactly the same moment in the series, A would run out of time again and terminate its search at exactly the same point.)

So what are we to say? Is our toy world really a world without prevention, without offense and defense, without lost opportunities, without the thrust and parry of genuine agency, without genuine possibilities? Admittedly, our chess programs, like insects or fish, are much too simple agents to be plausible candidates for morally significant free will, but we contend that the determinism of their

world does not rob them of their different powers, their different abilities to avail themselves of the opportunities presented. If we want to understand what is happening in that world, we may, indeed must, talk about how their choices cause their circumstances to change, and about what they *can* and *cannot* do.

Suppose we find two games in the series in which the first twelve moves are the same, but with A playing White in the first game and Black in the second. At move 13 in the first game, B "blunders" and its pattern goes downhill from there. At move 13 in the second game, A, in contrast, finds the saving move, castling, and goes on to win. "B *could have castled* at that point in the first game," says an onlooker, echoing Austin. True or false? The move, castling, was just as legal the first time, so in *that* sense, it was among the "options" available to B. Suppose we find, moreover, that castling was not only one of the represented candidate moves for B, but that B in fact undertook a perfunctory exploration of the consequences of castling, abandoned, alas, before its virtues were revealed. Could B have castled? What are we trying to find out? Looking at *precisely* the same case, again and again, is utterly uninformative, but looking at *similar* cases is in fact diagnostic. If we find that in many similar circumstances in other games, B *does* pursue the evaluation slightly farther, discovering the virtues of such moves and making them—if we find, in the minimal case, that flipping a single bit in the random number generator would result in B's castling—then we support ("with further experiments") the observer's conviction that B could have castled then. We would say, in fact, that B's failure to castle was a fluke, bad luck with the random number generator. If, on the contrary, we find that discovering the reasons for castling requires far too much analysis for B to execute in the time available (although A, being a stronger player, is up to the task), then we will have grounds for concluding that no, B, unlike A, could not have castled. To imagine B castling would require too many alterations of reality; we would be committing an error alluded to earlier, making X too large.

In sum, using the narrow method to choose X is useless if we want to explain the patterns that are manifest in the unfolding data. It is only if we "wiggle the events" (as David Lewis has said), looking *not* at "conditions as they precisely were" but at nearby neighboring worlds, that we achieve any understanding at all.[27] Once we expand X a little, we discover that B has additional options, in a sense both informative and morally relevant (when we address worlds beyond the chessboard). The burden rests with incompatibilists to explain why "real" possibility demands a narrow choice of X—or why we should be interested in such a concept of possibility, regardless of its "reality."

As we have seen, possibilities of the broader, more interesting variety can exist quite comfortably in deterministic worlds. Indeed, introducing indeterminism adds nothing in the way of worthwhile possibilities, opportunities, or competences to a universe. If in our sample deterministic world program A always beats program B, then replacing the pseudo-random number generator with a genuinely

indeterministic device will not help B at all: A will *still* win every time. Though pseudo-random generators may not produce genuinely random output, they come so close that no ordinary mortal can tell the difference. A superior algorithm like A's will hardly stumble when faced with so inconsequential a change. And analogous conclusions could well apply in meatier universes like ours. To put it graphically, the universe could be deterministic on even days of the month and indeterministic on odd days, and we would never notice a difference in human opportunities or powers; there would be just as many triumphs, and just as many lamentable lapses, on October 4 as on October 3 or October 5. (If your horoscope advised you to postpone any morally serious decision to an odd-numbered day, you would have no more reason to follow this advice than advice to wait for a waning moon.)

5. Some Related Fears

In passing we mention a number of other misguided worries about determinism, clustered about the basic fear of lost possibilities. Some thinkers have suggested that the truth of determinism might imply one or more of the following disheartening claims: all trends are permanent, character is by and large immutable, and it is unlikely that one will change one's ways, one's fortunes, or one's basic nature in the future. Ted Honderich,[28] for example, has maintained that determinism would somehow squelch what he calls our life hopes:

> If things have gone well for a person, there is more to hope for in what follows on the assumption that the entire run of his or her life is fixed. . . . If things have not gone well, or not so well as was hoped, it is at least not unreasonable to have greater hopes on the assumption that the whole of one's life is not fixed, but is connected with the activity of the self. . . . Given the sanguine premiss of our reasonableness, there is reason to think that we do *not* tend to the idea of a fixed personal future. (Honderich 1988; p. 388–89)

Clearly such anxieties originate in a vague sense that true possibilities (for an improved lot, say) disappear under determinism.

One readily sees the baselessness of such fears by referring again to the field of computer science. Programmers have already demonstrated how deterministic computer algorithms can adapt themselves to changes in the environment and learn from their mistakes.[29] Chess programs A and B from the previous section could well incorporate such talents. If initially mediocre B possesses these abilities and A does not, then we may ultimately find B emerging victorious. And if B has this sort of structure in a deterministic world, its enviable capacity will not im-

prove with the introduction of a genuinely indeterministic random-number generator. Nor will adding indeterminism to the universe help it if it lacks this ability.

In general, there is no paradox in the observation that certain phenomena are *determined* to be changeable, chaotic, and unpredictable, an obvious and important fact that philosophers have curiously ignored. Honderich finds disturbing the notion that we might have a "fixed personal *future*," but the implications of this notion are entirely distinct from the implications of having a "fixed personal *nature*." The latter is cause for dismay, perhaps, but not the former, for it could very well be one's fixed personal future to be blessed with a protean nature, highly responsive to the "activity of the self." The total set of personal futures, "fixed" or not, contains all sorts of agreeable scenarios, including victories over adversity, subjugations of weakness, reformations of character, even changes of luck. It could be just as determined a fact that you *can* teach an old dog new tricks as that you can't. The question to ask is, Are old dogs the kinds of things that can be taught new tricks? We rightly care about being the sorts of entities whose future trajectories are not certain to repeat the patterns found in the past. The general thesis of determinism has no implications about such issues—for answers to these questions, we must turn to specific fields like biology and social science (which themselves might be either deterministic or indeterministic domains).[30] And as the next section will show, creativity, the ability to author something of "originative value," is similarly independent of determinism.

6. Determinism and Causation (Thesis 2)

The hunch that determinism would eliminate some worthwhile type of causation from the universe has even less merit than the claim that it eliminates possibilities. We suspect this fear stems from the conflation of causal necessity with causal sufficiency—as we have seen, our language makes this confusion all too easy. Determinism is essentially a doctrine concerned with sufficiency: if σ_0 is a (mind-bogglingly complex) sentence that specifies in complete detail the state of the universe at t_0 and σ_1 similarly specifies the universe at a later time t_1, then determinism dictates that σ_0 is sufficient for σ_1 in all physically possible worlds. But determinism tells us nothing about what earlier conditions are *necessary* to produce σ_1, or any other sentence ψ for that matter. Hence, since causation generally presupposes necessity, the truth of determinism would have little bearing on the validity of our causal judgments.[31]

For example: according to determinism, the precise condition of the universe one second after the big bang (call the corresponding sentence σ_0) causally sufficed to produce the assassination of John F. Kennedy in 1963 (sentence ψ). Yet there is no reason at all to claim that σ_0 caused ψ. Though sufficient, σ_0 is hardly necessary. For all we know, Kennedy might well have been assassinated anyway, even if some different conditions had obtained back during the universe's birth.[32] More plausible causes of the event would include "A bullet followed a course directed at Kennedy's body"; "Lee Harvey Oswald pulled the trigger on his gun"; perhaps "Kennedy was born"; conceivably "Oswald was born."[33] But conspicuously absent from this list are microscopically detailed descriptions of the universe billions of years prior to the incident. Incompatibilists who assert that under determinism σ_0 "causes" or "explains" ψ miss the main point of causal inquiry.

In fact, determinism is perfectly compatible with the notion that some events have no cause at all. Consider the sentence "The devaluation of the rupiah caused the Dow Jones average to fall." We rightly treat such a declaration with suspicion; are we really so sure that among nearby universes the Dow Jones fell *only* in those where the rupiah fell first? Do we even imagine that every universe where the rupiah fell experienced a stock market sell-off? Might there not have been a confluence of dozens of factors that jointly sufficed to send the market tumbling but none of which by itself was essential? On some days, perhaps, Wall Street's behavior has a ready explanation; yet at least as often we suspect that no particular cause is at work. And surely our opinions about the market's activities would remain the same, whether we happened to adopt Newton's physics or Schrödinger's.

Of course, one might wonder why it is that causal necessity matters to us as much as it does. Let us return for a moment to chess programs A and B. Suppose our attention is drawn to a rare game in which B wins, and we want to know "the cause" of this striking victory. The trivial claim that B's win was "caused" by the initial state of the computer is totally uninformative. Of course the total state of the toy universe at prior moments was *sufficient* for the occurrence of the win; we want to know which features were *necessary,* and thereby understand what such rare events have in common. We want to discover those features, the absence of which would most directly be followed by B's loss, the default outcome. Perhaps we will find a heretofore unsuspected flaw in A's control structure, a bug that has only just now surfaced. Or perhaps the victory is a huge coincidence of conditions that require no repair, since the probability of their recurrence is effectively zero. Or we might find an idiosyncratic island of brilliance in B's competence, which once diagnosed would enable us to say just what circumstances in the future might permit another such victory for B.

Rationality *requires* that we evaluate necessary conditions at least as carefully as sufficient conditions. Consider a man falling down an elevator shaft. Although he doesn't know exactly which possible world he in fact occupies, he does know

one thing: he is in a set of worlds *all* of which have him landing shortly at the bottom of the shaft. Gravity will see to that. Landing is, then, *inevitable* (un-avoid-able) because it happens in every world consistently with what he knows. But perhaps *dying* is not inevitable. Perhaps in some of the worlds in which he lands, he survives. Those worlds do not include any in which he lands headfirst or spreadeagled, say, but there may be worlds in which he lands in a toes-first crouch and lives. There is some elbow room. He can rationally plan action on the assumption that living is possible, and even if he cannot discover sufficient conditions to guarantee survival, he may at least improve the odds by taking whatever actions are necessary.[34]

In closing, let us return to the human desire pinpointed by Kane that motivates so much of this debate: the desire to be able to take full credit as the creators and causes of change in the world. Consider for instance the wish that we (Taylor and Dennett) have to be acknowledged as the authors of this essay. Suppose that determinism turns out to be true. Would that in any way undercut our claim that our activity nevertheless played an essential role in this essay's creation? Not in the least, even after we factor in the earlier deeds of our parents and teachers. Without our efforts, it is safe to say that no essay exactly like this (or even closely similar) would have been produced.[35] Hence we are entitled to claim some "originative value" for our unique accomplishment. The thirst for originality and causal relevance is not to be quenched by abstruse quantum events: all that we require is the knowledge that without our presence, the universe would have turned out significantly different.

Appendix: Van Inwagen's Consequence Argument

Peter van Inwagen (1975) hopes to bolster the incompatibilist sense of lost causal powers with the following basic argument:

1. Let φ be some event that actually occurs in agent A's life (missing a putt, say). Also let σ_0 be a comprehensive description of the universe's state at some time in the remote past, and let λ be a statement of the laws of nature.
2. Then, assuming determinism, $\lambda \wedge \sigma_0 \Rightarrow \varphi$ applies in every possible world. Equivalently, $\sim\varphi \Rightarrow \sim (\lambda \wedge \sigma_0)$.
3. If A has the power to cause α and $\alpha \Rightarrow \beta$ obtains in every possible world, then A has the power to cause β.

4. So if A has the power to cause $\sim\varphi$, then A has the power to cause the falsity of either λ or σ_0, which is absurd.
5. Therefore A lacks the power to cause $\sim\varphi$.

This argument illustrates nicely the confusion that causal necessity and sufficiency engender. As we have argued, counterfactual necessity is the single most crucial condition for causation, and accordingly we would recommend that van Inwagen's "power to cause α" be rendered as follows:

> A has the *power to cause* α iff for some sentence γ describing an action of A and a world f close to actuality, $\gamma \wedge \alpha$ holds in f and $\alpha \Rightarrow \gamma$ in every world similar to f.

In other words, within some cluster of nearby worlds, there is a possible action of A (called γ) that is a necessary condition for α to occur. But under this definition, line 3 has no warrant whatever. Line 3 hypothesizes that $\alpha \Rightarrow \gamma$ in a cluster of nearby worlds, and that $\alpha \Rightarrow \beta$ in every world; if we could deduce that $\beta \Rightarrow \gamma$ in this cluster, we would be home free. But of course in Logic 101 we learn that $\alpha \Rightarrow \gamma$ and $\alpha \Rightarrow \beta$ do not entail $\beta \Rightarrow \gamma$, and so line 3 fails, and van Inwagen's argument with it.

NOTES

1. Nozick (1981: 313), "We want it to be true that in that very same situation we could have done (significantly) otherwise, so that our actions will have originative value."
2. Austin (1961: 166).
3. Kane (1996. 100).
4. Ibid. (p. 4).
5. Ibid. (p. 33).
6. Ibid. (p. 35).
7. Ginet (1990: 9).
8. See Quine (1980) for a discussion of these pitfalls.
9. Idem (1969: 147–55).
10. The average educated person's casual working assumptions about the cosmos still resemble the Democritean account, and philosophers traditionally rely on nothing more sophisticated when exploring the implications of determinism and indeterminism, causation and possibility.

Our suggestion that possible worlds simply *are* functions of the appropriate form may seem disturbingly reductive, particularly when one contemplates the particular function(s) that correspond to the actual world; accordingly David Lewis takes pains

to distinguish possible worlds from their mathematical "handles." However one wishes to address these ontological scruples, nothing in the following discussion hinges on them.

11. Since we are restricting ourselves to the scientifically old-fashioned Democritean worlds, we would have trouble specifying the contents of Φ precisely—and besides, of course, we do not yet *know* all the laws of nature—but we can pretend that we know, and hence we can pretend that in most cases one can judge whether or not a particular world f accords with natural law.

John Horton Conway's Game of Life can be viewed as a particularly simple pseudo-Democritean universe, eliminating one spatial dimension and quantizing time. (See Dennett 1991: 27–51 or Idem 1995, for an introduction to Life.) The set of all possible sequences of bitmaps is then Ω, and the single (deterministic) rule of Life "physics" applied to every "initial" state gives us the subset Φ of Ω. Every variation on Conway's "physics" generates a different subset Φ.

12. Of course, these predicates unleash a horde of problems concerning vagueness, subjectivity, and (in such cases as "believes that snow is white") intentionality, but difficulties along these lines do not imperil the basic approach.

13. Lewis (1973a), *passim.*

14. See, for example, (Tooley 1987).

15. A vast amount of ink his been spilled arguing that the direction of causation is either independent of or logically prior to the direction of time, and to address the matter here would require too lengthy a digression. So we merely note the issue and tentatively take the direction of time as a given (originating ultimately in the Second Law of Thermodynamics) from which the direction of causation derives.

Gasking (1955) raises a number of interesting cases in which cause and effect appear to be simultaneous: for instance, if a piece of iron attains a temperature of (say) 1000°C and thereupon starts to glow, we still distinguish the former as cause and the latter as effect. But this apparent exception to the rule has a ready explanation that Gasking himself hints at: when a speaker refers to the iron "reaching 1000°," she is envisioning this event as the endpoint in a lengthy heating process. The heating process *does* precede the glowing, and so the latter is considered an effect.

Another category of "exceptions" includes diseases and their symptoms (say, a cold and sneezing), which might sometimes arise simultaneously. Yet often enough diseases *do* precede their symptoms, while symptoms (by definition) never appear before their diseases. Accordingly we grant diseases the status of "cause."

16. Notice that we do not in the previous clauses make any provision to ensure the transitivity of causation. Lewis (2000: 191–95), among others, feels it important to guarantee transitivity by making "causation" the ancestral of "causal dependence." But Lewis himself provides many examples of transitivity's counterintuitive consequences. For instance, suppose that agent A wants to travel to New York. Agent B, hoping to thwart A, lets the air out of the tires on A's car. In consequence, A takes the train instead and reaches New York only slightly behind schedule. If causation is transitive, than B has "caused" A's successful arrival, despite the fact that the two sentences "B lets the air out of A's tires" and "A arrives in New York" satisfy none of our more crucial conditions. Lewis finds the awkward implications of transitivity acceptable; we remain unpersuaded.

Hall (2000) goes to even greater lengths to defend transitivity. His account seems to imply that a pebble on the train tracks south of Paris that minutely alters the course of

the Orient Express is a "cause" of the train's arrival in Istanbul several days later. Paul's "Aspect Causation" (2000) suggests a possible diagnosis for Hall's willingness to countenance such bizarre conclusions, as stemming from an overeager acceptance of the premise that causation is a relation between "events" (however this problematic term may be defined). At any rate, notice that on our account one can consistently consider false the sentence "Pebble *p*'s lying on the tracks south of Paris caused the train's arrival in Istanbul," while accepting "Pebble *p*'s lying on the tracks south of Paris caused the train's arrival in Istanbul via a minutely altered course in France."

17. Lewis (2000).

18. Obviously, a sentence like "Drugs or aliens caused Elvis's premature demise" abbreviates the cumbersome "Drugs caused Elvis's premature demise or aliens caused Elvis's premature demise"—a disjunction of two separate causes, not a single disjunctive cause.

19. Invoking causal sufficiency in this way solves, to our satisfaction, all of the analogous problem cases raised by Shaffer (2000). Note that Shaffer rather misleadingly suggests that "counterfactual accounts of causation" must always be formulated solely in terms of necessity (ibid.: 176). We, on the contrary, consider our account essentially "counterfactual" even though it allows for sufficiency along with necessity.

Lewis's formulation (Lewis 2000) of "Causation as Influence" can be viewed as an indirect way of introducing sufficiency into an originally necessity centered account. For present purposes we consider our approach more illuminating, but both strategies point in the same general direction.

20. A doubly elaborated version of the example due originally to McLaughlin (1925), first elaborated in Hart and Honoré (1959). The Hart and Honoré version has one less twist: "Suppose A is entering a desert. B secretly puts a fatal dose of poison in A's water keg. A takes the keg into the desert where C steals it; both A and C think it contains water. A dies of thirst. Who kills him?"

21. When Austin speaks of further experiments, could he be referring to experiments in the high-tech labs of physicists and microbiologists, experiments that would convince him that his brain amplifies indeterministic quantum events? Given the extreme impracticality of such experiments, and Austin's overall skepticism about the relevance of science in these contexts ("[A modern belief in science] is not in line with the traditional beliefs enshrined in the word *can*," Austin 1961: 166), this interpretation seems unlikely. But this is precisely the direction in which Kane and some other incompatibilists have headed. See also Dennett (1984: 133–37).

22. We are restricting our attention to programs that do not require or accept input from the external world, which could, of course, be random in any of several senses. The easiest way to ensure that there is variation in subsequent runs of a program is to have it call for inputs of these sorts: the time taken from the computer's clock, the presence or absence of a pulse from a Geiger counter, the last digit in the latest Dow Jones Industrial Average as taken off the Internet, and so on.

23. All this is independent of whether or not either chess program can "learn from its experience," which is another way their internal state could change over time to guarantee that no two games were the same.

24. Another case in which we could know *all* the deterministic microdetails but be baffled about how to explain the causal regularities is Dennett's example of the two black boxes (1995: 412–22).

25. Dennett (1978: 107).

26. Compare the comet plunging toward earth that is intercepted at the last minute by the other comet, unnoticed till then, that had been on its collision trajectory since its birth millions of years ago (Dennett 1984: 124).

27. If we exclude such variation, then trivially, castling in the second game was not "open to B," to use Ginet's terminology. Recall that Ginet requires that "nothing that exists up to that moment stands in the way of my doing next any one of the alternatives." The narrow method has the effect of treating the precise state of B's contemplation of the option of castling as something *external,* as something that can itself "stand in the way" at the moment of choosing, guaranteeing that *nothing about B* could *explain* B's choice, whatever it is. As Dennett notes, "If you make yourself really small, you can externalize virtually everything" (1984: 143).

28. Honderich (1988).

29. They have also demonstrated, all too often, the possibility of programs losing competence over time by accumulating deleterious effects from bugs. At any rate, just how significant are the many examples of "machine learning" that have been produced to date? The answer is contested, and it is true that the best chess programs today do *not* include substantial "unsupervised" learning capacities. Still, the feasibility of genuine learning in computer programs has not been in doubt since the self-improving checkers program created by Arthur Samuel in the 1950s. (See Dennett 1995: 207–12 for details.) John McCarthy has posed the question of what the minimal life-world configuration is, in which occupants learn the physics of their own world (ibid.: 175). One might also ask, Which variations on Conway's physics generate possible worlds in which occupants can know or learn anything at all?

30. This paragraph is drawn, with revisions, from Dennett 1988.

31. See the appendix to this chapter for an additional example of the conflation of necessity and sufficiency (in van Inwagen's Consequence Argument).

32. Imagine that we take a snapshot of the universe at the moment of Kennedy's assassination, then alter the picture in some trivial way (by moving Kennedy 1 mm to the left, say). Then, following the (deterministic) laws of physics in reverse, we can generate a movie running all the way back to the Big Bang, obtaining a world in which σ_0 subtly fails.

33. Of course, the last two options fail the sufficiency test so badly that we prefer not to countenance them as causes. As explained earlier, sufficiency does have *some* relevance in assigning causes, but not the overwhelming importance that incompatibilists imply.

34. The dependence of this concept of possibility on *epistemic* considerations has been suggested before (see Dennett 1984: 147ff.) but mischaracterized. It is true that if determinism held, and if the man knew *exactly* which world he inhabited, he would already know his fate.

35. Similarly, Deep Blue, in spite of its being a deterministic automaton, authored the games of chess that vanquished Kasparov. No one *else* was their author; Murray Campbell and the IBM team that created Deep Blue cannot claim credit for those games; *they* did not see the moves. The vast exploratory activity of Deep Blue itself was the originating cause of those magnificent games.

THE OXFORD HANDBOOK OF

FREE WILL

Edited by

ROBERT KANE

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford New York
Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Taipei Tokyo Toronto

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

First issued as an Oxford University Press paperback, 2005

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
The Oxford handbook of free will / edited by Robert Kane.
p. cm.
Includes bibliographical references and index.
ISBN- 13 978-0-19-517854-8

1. Free will and determinism. 2. Philosophy, Modern—20th century.
3. Ethics, Modern—20th century. I. Kane, Robert, 1938–
BJ1461 .F74 2002
123'.5—dc21 00–052872

7 9 8 6

Printed in the United States of America
on acid-free paper