Daniel Dennett, *Elbow Room*, Ch. 6.

"Could Have Done Otherwise"

1. Do We Care Whether We Could Have Done Otherwise?

In the midst of all the discord and disagreement among philosophers about free will, there are a few calm islands of near unanimity. As van Inwagen notes:

> Almost all philosophers agree that a necessary condition for holding an agent responsible for an act is believing that the agent *could have* refrained from performing that act. (van Inwagen 1975, p.189)

But if this is so, then whatever else I may have done in the preceding chapters, I have not yet touched the central issue of free will, for I have not yet declared a position on the "could have done otherwise" principle: the principle that holds that one has acted freely (and responsibly) only if one could have done otherwise. It is time, at last, to turn to this central, stable area in the logical geography of the free will problem. First I will show that this widely accepted principle is simply false. Then I will turn to some residual problems about the meaning of "can"—Austin's frog at the bottom of the beer mug (see chapter one, page 19).

The "could have done otherwise" principle has been debated for generations, and the favorite strategy of compatibilists—who must show that free will and determinism are compatible after all—is to maintain that "could have done otherwise" does not mean what it seems at first to mean; the sense of the phrase denied by determinism is irrelevant to the sense required for freedom. It is so obvious that this is what the compatibilists have to say that many skeptics view the proffered compatibilist "analyses" of the meaning of "could have done otherwise" as little more than self-deceived special pleading. James (1921, p.149) called this theme "a quagmire of evasion" and Kant (*Critique of Practical Reason*, Abbot translation 1873, p.96) called it a "wretched subterfuge."

Instead of rising to the defense of any of the earlier analyses—many of which are quite defensible so far as I can see—I will go on the offensive. I will argue that *whatever* "could have done otherwise" actually means, it is not what we are interested in when we care about whether some act was freely and responsibly performed. There is, as van Inwagen notes, something of a tradition of simply assuming that the intuitions favoring the "could have done otherwise" principle are secure. But philosophers who do assume this do so in spite of fairly obvious and familiar grounds for doubt.

One of the few philosophers to challenge it is Frankfurt, who has invented a highly productive intuition pump that generates counterexamples in many flavors: cases of overdetermination, where an agent deliberately and knowingly chose to do something, but where—thanks typically to some hovering bogeyman—if he hadn't so chosen, the bogeyman would have seen to it that he did the thing anyway (Frankfurt 1969, but see also van Inwagen 1978 and 1983, and Fischer 1982). Here is the basic, stripped-down intuition pump (minus the bells and whistles on the variations, which will not concern us—but only because we will not be relying on them):

> Jones hates Smith and decides, in full possession of his faculties, to murder him. Meanwhile Black, the nefarious neurosurgeon (remember him?), who also wants Smith dead, has implanted something in Jones' brain so that *just in case Jones changes his mind* (and chickens out), Black, by pushing his special button, can put Jones back on his murderous track. In the event Black doesn't have to intervene; Jones does the deed all on his own.

In such a case, Frankfurt claims, the person would be responsible for his deed, since he chose it with all due deliberation and wholeheartedness, in spite of the lurking presence of the overdeterminer whose hidden presence makes it the case that Jones couldn't have done otherwise.

I accept Frankfurt's analysis of these cases (that is, I think they can be defended against the objections raised by van Inwagen, Fischer, and others), and think these thought experiments are useful in spite of their invocation of imaginary bogeymen, for they draw attention to the importance, for responsibility, of the actual causal chain of deliberation and choice running through the agent—whatever may be happening elsewhere.

But Frankfurt's strategy seems to me to be insufficiently ambitious. Although he takes his counterexamples to show that the "could have done otherwise" principle—which he calls the principle of alternate possibilities—is irremediably false, his counterexamples are rather special and unlikely cases, and they invite the defender of the principle to try for a patch: modify the principle slightly to take care of Frankfurt's troublesome cases. Exotic circumstances do little or nothing to dispel the illusion that in the normal run of things, where such overdetermination is lacking, the regnant principle is indeed that if a person could not have refrained (could not have done otherwise), he would not be held responsible. But in fact, I will argue, it is seldom that we even *seem* to care whether or not a person could have done otherwise. And when we do, it is often because we wish to draw the opposite conclusion about responsibility from the one tradition endorses.

"Here I stand," Luther said. "I can do no other." Luther claimed that he could do no other, that his conscience made it *impossible* for him to recant. He might, of course, have been wrong, or have been deliberately overstating the truth. But even if he was—perhaps especially if he was—his declaration is testimony to the fact that we simply do not exempt someone from blame or praise for an act because we think he could do no other. Whatever Luther was doing, he was not trying to duck responsibility.

There are cases where the claim "I can do no other" is an avowal of frailty: suppose what I ought to do is get on the plane and fly to safety, but I stand rooted to the ground and confess that I can do no other—because of my irrational and debilitating fear of flying. In such a case I can do no other, I claim, because my rational control faculty is impaired. But in other cases, like Luther's, when I say I cannot do otherwise I mean that I cannot because I see so clearly what the situation is and because my rational control faculty is *not* impaired. It is too obvious what to do; reason dictates it; I would have to be mad to do otherwise, and since I happen not to be mad, I cannot so otherwise. (Notice, by the way, that we say it was "up to" Luther whether or not to recant, and we do not feel tempted to rescind that judgment when we learn that he claimed he could do no other. Notice, too, that we often say things like this: "If it were up to me, I know for certain what I would do.")

I hope it is true—and think it very likely is true—that it would be impossible to induce me to torture an innocent person by offering me a thousand dollars. "Ah"—comes the objection—"but what if some evil space pirates were holding the whole world ransom, and promised not to destroy the world if only you would torture an innocent person? Would that be something you would find impossible to do?" Probably not, but so what? That is a vastly different case. If what one is interested in is whether *under the specified circumstances* I could have done otherwise, then the other case mentioned is utterly irrelevant. I claimed it would not be possible to induce me to torture someone *for a thousand dollars*. Those who hold dear the principle of "could have done otherwise" are always insisting that we should look at whether one could have done otherwise in *exactly* the same circumstances. I claim something stronger; I claim that I could not do otherwise even in any roughly similar case. I would *never* agree to torture an innocent person for a thousand dollars. It would make no difference, I claim, what tone of voice the briber used, or whether or not I was tired and hungry, or whether the proposed victim was well illuminated or partially concealed in shadow. I am, I hope, immune to all such offers.

Now why would anyone's intuitions suggest that if I am right, then if and when I ever have occasion to refuse such an offer, my refusal would not count as a responsible act? Perhaps this is what some people think: they think that if I were right when I claimed I could not do otherwise in such cases, I would be some sort of zombie, "programmed" always to refuse thousand-dollar bribes. A genuinely free agent, they think, must be more volatile somehow. If I am to be able to listen to reason, if I am to be flexible in the right way, they think, I mustn't be too dogmatic. Even in the most preposterous cases, then, I must be able to see that "there are two sides to every question." I must be able to pause, and weigh up the pros and cons of this suggested bit of lucrative torture. But the only way I could be constituted so that I can always "see both sides"—no matter how preposterous one side is—is by being constituted so that in any particular case "I could have done otherwise."

That would be fallacious reasoning. Seeing both sides of the question does not require that one not be overwhelmingly persuaded, in the end, by one side. The flexibility we want a responsible agent to have is the flexibility to recognize the one-in-a-zillion case in which, thanks to that thousand dollars, not otherwise obtainable, the world can be saved (or whatever). But the general capacity to respond flexibly in such cases does not at all require that one "could have done otherwise" in the particular case, but only that under – some variations in the circumstances—the variations that matter—one would do otherwise.

It might be useful to compare two cases that seem quite different at first, but belong on a continuum.

I. Suppose I know that if I ever see the full moon, I will probably run amok and murder the first person I see. So I make careful arrangements to have myself locked up in a windowless room on several nights each month. I am thus rendered unable to do the awful things I would do otherwise. Moreover, it is thanks to my own responsible efforts that I have become unable to do these things. A fanciful case, no doubt, but consider the next case, which is somewhat more realistic.

II. Suppose I know that if I ever see a voluptuous woman walking unescorted in a deserted place I will probably be overcome by lust and rape her. So I educate myself about the horrors of rape from the woman's point of view, and enliven my sense of the brutality of the crime so dramatically that if I happen to encounter such a woman in such straits, I am unable to do the awful thing I would have done otherwise. (What may convince me that I would otherwise have done this thing is that when the occasion arises I experience a considerable inner tumult; I discover myself shaking the bars of the cage I have built for myself.) Thanks to my earlier responsible efforts, I have become quite immune to this rather more common sort of possession; I have done what had to be done to render certain courses of action *unthinkable* to me. Like Luther, I now can do no other.

Suppose—to get back all the way to realism—that our parents and teachers know that if we grow up without a moral education, we will become selfish, untrustworthy and possibly dangerous people. So they arrange to educate us; and thanks to their responsible efforts, our minds recoil from thoughts of larceny, treachery and violence. We find such alternatives unthinkable under most normal circumstances, and moreover have been taught to think ahead for ourselves and to contribute to our own moral development. Doesn't a considerable part of being a responsible person consist in making oneself unable to do the things one would be blamed for doing if one did them? Philosophers have often noted, uneasily, that the difficult moral problem cases, the decisions that "might go either way," are not the only, or even the most frequent, sorts of decisions for which we hold people responsible. They have seldom taken the hint to heart, however, and asked whether the "could have done otherwise" principle was simply wrong.

I grant that we do indeed often ask ourselves whether an agent could have done otherwise—and in particular whether or not we ourselves could have done otherwise—in the wake of some regrettable act. But we never show any interest in trying to answer the question we have presumably just asked! Defenders of the principle suppose that there is a sense of "could have done otherwise" according to which, if determinism is true, no one ever could have done otherwise than he did. Suppose they are right that there is such a sense. Is it the sense we intend when we use the words "could he have done otherwise?" to inaugurate an inquiry into an agent's responsibility for an act he committed? It is not. In pursuing such inquiries we manifestly ignore the sort of investigations that would have to be pursued if we really were interested in the answer to that question, the metaphysicians' question about whether or not the agent was completely determined by the state of the universe at that instant to perform that action.

If our responsibility really did hinge, as this major philosophical tradition insists, on the question of whether we ever could do otherwise than we in fact do *in exactly those circumstances*, we would be faced with a most peculiar problem of ignorance: it would be unlikely in the extreme, given what now seems to be the case in physics, that anyone would ever know whether anyone has ever been responsible. For today's orthodoxy is that indeterminism reigns at the subatomic level of quantum mechanics, so in the absence of any general and accepted argument for universal determinism, it is possible for all we know that our decisions and actions are truly the magnified, macroscopic effects of quantum-level indeterminacies occurring in our brains. But it is also possible, for all we know, that even though indeterminism reigns in

our brains at the subatomic quantum mechanical level, our macroscopic decisions and acts are all themselves determined; the quantum effects could just as well be self-canceling, not amplified (as if by organic Geiger counters in the neurons). And it is extremely unlikely, given the complexity of the brain at even the molecular level (a complexity for which the word "astronomical" is a vast understatement), that we could ever develop good evidence that any particular act was such a large-scale effect of a critical subatomic indeterminacy. So if someone's responsibility for an act did hinge on whether, at the moment of decision, that decision was (already) determined by a prior state of the world, then barring a triumphant return of universal determinism in microphysics (which would rule out all responsibility on this view), the odds are very heavy that we will never have *any* reason to believe of any particular act that it was or was not responsible. The critical difference would be utterly inscrutable from every macroscopic vantage point, and practically inscrutable from the most sophisticated microphysical vantage point imaginable.

Some philosophers might take comfort in this conclusion, but I would guess that *only* a philosopher could take comfort in it. To say the very least it is hard to take seriously the idea that something that could matter so much could be so magnificently beyond our ken. (Or look at the point another way: those who claim to know that they have performed acts such that they could have done otherwise in exactly those circumstances must admit that they proclaim this presumably empirical fact without benefit of the slightest shred of evidence, and without the faintest hope of ever obtaining any such evidence.)[1]

[1] Raab (1955) claims that the metaphysical question about "the absence of causality" is "untestable," and notes the peculiarity of taking such an unanswerable question seriously. Raab's reason for declaring such questions unanswerable rests on the claim—true, no doubt—that no agent has any privileged access to whether or not his action was caused. All this shows is that such questions ought not to be addressed exclusively to the agent. My point is that *no* investigation could shed any reliable light on this.

Given the sheer impossibility of conducting any meaningful investigation into the question of whether or not an agent could have done otherwise, what can people think they are doing when they ask that question in particular cases? They must take themselves to be asking some other question. They are right; they are asking a much better question. (If a few people have been asking the unanswerable metaphysical question, they were deluded into it by philosophy.) The question people are really interested in asking is a better question for two reasons: it is usually empirically answerable, and its answer matters. For not only is the traditional metaphysical question unanswerable; its answer, even if you knew it, would be useless.

What good would it do to know, about a particular agent, that on some occasion (or on every occasion) he could have done otherwise than he did? Or that he could not have done otherwise than he did? Let us take the latter case first. Suppose you knew (because God told you, presumably) that when Jones pulled the trigger and murdered his wife at time *t*, he could *not* have done otherwise. That is, given Jones' microstate at *t* and the complete microstate of Jones' environment (including the gravitational effects of distant stars, and so on) at *t*, no other Jones-trajectory was possible than the trajectory he took. If Jones were ever put back into exactly that state again, in exactly that circumstance, he would pull the trigger again. And if he were put in that state a million times, he would pull the trigger a million times.

Now if you learned this, would you have learned anything about Jones? Would you have learned anything about his character, for instance, or his likely behavior on merely similar occasions? No. Although people are physical objects which, like atoms or ball bearings or bridges, obey the laws of physics, they are not only more complicated than anything else we know in the universe, they are also designed to be so sensitive to the passing show that they never can be in the same microstate twice. One doesn't even have to descend to the atomic level to establish this. People learn, and remember, and get bored, and shift their attention, and change their interests so incessantly, that it is as good as infinitely unlikely that any person is ever in the same (gross) *psychological* or *cognitive* state on two occasions. And this would be true even if we engineered the surrounding environment to be "utterly the same" on different occasions—if only because the second time around the agent would no doubt think something that went unthought the first time, like "Oh my, this all seems so utterly familiar; now what did I do last time?" (see chapter two, page 33)

There is some point in determining how a bridge is caused to react to some very accurately specified circumstances, since those may be circumstances it will actually encounter *in its present state* on a future occasion. But there would be no payoff in understanding to be gained by determining the micro-causation of the behavior of a human being in some particular circumstance, since he will certainly never confront that micro-circumstance again, and even if he did, he would certainly be in a significantly different reactive state at the time.

Learning (from God, again) that a particular agent was *not* thus determined to act would be learning something equally idle, from the point of view of character assessment or planning for the future. As we saw in chapter five, the undetermined agent will be no more flexible, no more versatile, no more sensitive to nuances, no more reformable, than his deterministic cousin.

So if anyone is interested at all in the question of whether or not one could have done otherwise in exactly the same circumstances (and internal state), this will have to be a particularly pure metaphysical curiosity— that is to say, a curiosity so pure as to be utterly lacking in any ulterior motive, since the answer could not conceivably make any noticeable difference to the way the world went.[2]

[2] Nozick (1981, p. 313) claims that we all want "originative value," but the only conditions under which we would have this are (on his analysis) conditions that

Why, though, does it still seem as if there ought to be a vast difference, somehow visible from the ordinary human vantage point, between a world in which we could not do otherwise and a world in which we could? Why should determinism still seem so appalling? Perhaps we are misled by the God's-eye-view image, "*sub specie aeternitatis*," in which we spy our own life-trajectories in space and time laid out from birth to death in a single, fixed, rigid, unbranching, four-dimensional "space-time worm," pinned to the causal fabric and unable to move. (Causation, in Hume's fine metaphor, is "the cement of the universe" (Mackie 1974), so perhaps we see our entire lives as *cast in concrete*, trapped like a fossil in the unchanging slab of space-time.)

What we would like, it seems, is for someone to show us that we can *move about* in that medium. But this is a confusion; if we feel this yearning it is because we have forgotten that time is one of the dimensions we have spatialized in our image. Scanning from left to right is scanning from past to future, and a vertical slice of our image captures a single moment in time. To have elbow room in that medium—to be able to wiggle and squirm in between the fixed points of birth and death for instance—would not be to have the power to choose in an undetermined way, but to have the power to choose two or more courses *at one time*.

Is that what we want—to have our cake and eat it too? To have chosen *both* to marry and to remain unmarried, *both* to pull the trigger and to drop the gun? If that is the variety of free will we want, then whether or not it might be worth wanting, we can be quite confident that it must elude us—unless, perhaps, we adopt Everett's many-worlds interpretation of quantum mechanics, in which case it just might follow that we do lead a zillion lives (though our many alter egos, alas, could never get together and compare notes)!

If we let go of that fantasy and ask what we really, soberly want, we find a more modest hope: while there are indeed times when we would give anything to be able to go back and undo something in the past, we recognize that the past is closed for us, and we would gladly settle for an "open future." But what would an open future be? A future in which our deliberation is effective: a future in which if I decide to do A then I will do A, and if I decide to do B then I will do B; a future in which—since only one future is possible—the only possible thing that can happen is the thing I decide in the end to do.

2. *What We Care About*

If it is unlikely then that it matters whether or not a person could have done otherwise (when we look microscopically closely at the causation involved) what is the other question that we are really interested in when we ask "but could he have done otherwise?"

Once more I am going to use the tactic of first answering a simpler question about simpler entities. Consider a similar question that might arise about our deterministic robot, the Mark I Deterministic Deliberator. By hypothesis, it lives its entire life as a deterministic machine on a deterministic planet, so that whatever it does, it could not have done otherwise, if we mean that in the strict and metaphysical sense of those words that philosophers have concentrated on. Suppose then that one fine Martian day it makes a regrettable mistake; it concocts and executes a scheme that destroys something valuable—another robot, perhaps. I am not

apparently demand the metaphysical reading of "could have done otherwise": "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." Once again, is it plausible at all that something we care so much about (if Nozick is right) is something we could never know to be the case? Put another way, if originative value requires this, why would anyone care about having originative value?

supposing, for the moment, that it can regret anything,[3] but just that its designers, back on Earth, regret what it has done, and find themselves wondering a wonder that might naturally be expressed: *could it have done otherwise*?

They know it is a deterministic system, of course, so they know better than to ask the metaphysical question. Their question concerns the design of the robot; for in the wake of this regrettable event they may wish to redesign it slightly, to make this sort of event less likely in the future.[4] What they want to know, of course, is what information the robot was relying on, what reasoning or planning it did, and whether it did "enough" of the right sort of reasoning or planning. Of course in one sense of "enough" they know the robot did not do enough of the right sort of thing; if it had, it would have done the right thing. But it may be that the robot's design in this case could not really be improved. For it may be that it was making optimal use of optimally designed heuristic procedures—but this time, unluckily, the heuristic chances it took didn't pay off. Put the robot in a *similar* situation in the future, and thanks to no more than the fact that its pseudo-random number generator is in a different state, it will do something different; in fact it will usually do the right thing. It is tempting to add: it *could* have done the right thing on this occasion—meaning by this that it was well enough designed, at that time, to have done the right thing (its "character" is not impugned). Its failure depended on nothing but the fact that something *undesigned* (and unanticipatable) happened to intervene in the process in a way that made an unfortunate difference.

[3] Just because, for *this* purpose, I can consider a relatively simple robot. A robot that was self-made in the manner of the self-made selves of chapter four would be capable (I would claim) of regret.

[4] "We are scarcely ever interested in the performance of a communication-engineering machine for a single input. To function adequately it must give a satisfactory performance for a whole class of inputs, and this means a statistically satisfactory performance for the class of inputs which it is statistically expected to receive." (Wiener 1948, p. 55)

A heuristic program is not guaranteed to yield the "right" or sought-after result. Some heuristic programs are better than others; when one fails, it may be possible to diagnose the failure as assignable to some characteristic weakness in its design. But even the best are not foolproof, and when they fail, as they sometimes must, there may be no reason at all for the failure: as Cole Porter would say, it was just one of those things.

Such failures are not the only cases of failures that will "count" for the designers as cases where the system "could have done otherwise." If they discover that the robot's failure, on this occasion, was due to a "freak" bit of dust that somehow drifted into a place where it could disrupt the system, they may decide that this was such an unlikely event that there is no call to redesign the system to guard against its recurrence.[5] They will note that, in the micro-particular case, their robot could not have done otherwise; moreover, if (by remotest possibility) it ever found itself in exactly the same circumstance again, it would fail again.

But the designers will realize that they have no rational interest in doing anything to improve the design of the robot. It failed on the occasion, but its design is nevertheless above reproach. There is a difference between being optimally designed and being infallible. (See chapter seven.)

Consider yet another sort of case. The robot has a ray gun that it fires with 99.9 percent accuracy: That is to say, sometimes, over long distances, it fails to hit the target it was aiming at. Whenever it misses, the engineers want to know something about the miss: was it due to some systematic error in the controls, some foible or flaw that will keep coming up, or was it just one of those things—one of those "acts of God" in which, in spite of an irreproachable

[5] Strictly speaking, the recurrence of an event of this general type; there is no need to guard against the recurrence of the particular event (something logically impossible), or against the recurrence of an event of exactly the same type (something nomologically impossible).

execution of an optimally designed aiming routine, the thing just narrowly missed? There will always be such cases; the goal is to keep them to a minimum—consistent with cost- effectiveness of course. Beyond a certain point, it isn't worth caring about errors. Quine (1960, pp. 182 and 259) notes that engineers have a concept of more than passing philosophical interest: the concept of "don't-cares"—the cases that one is rational to ignore. When they are satisfied that a particular miss was a don't-care, they may shrug and say: "Well, it could have been a hit."

What concerns the engineers when they encounter misperformance in their robot is whether or not the misperformance is a telling one: does it reveal something about a pattern of systematic weakness, likely to recur, or an inappropriate and inauspicious linking between sorts of circumstances and sorts of reactions? Is this *sort* of thing apt to happen again, or was it due to the coincidental convergence of fundamentally independent factors, highly unlikely to recur? To get evidence about this they ignore the micro-details, which will never be the same again in any case, and just average over them, analyzing the robot into a finite array of macroscopically defined states, organized in such a way that there are links between the various degrees of freedom of the system. The question they then ask is this: are the links the right links for the task?[6]

This rationale for ignoring micro-determinism (wherever it may "in principle" exist) and squinting just enough to blur such fine distinctions into probabilistically related states and regions that can be *treated* as homogeneous is clear, secure, and unproblematic in science, particularly in engineering and biology, as we have seen. (See Wiener 1948 and Wimsatt 1980.) That does not mean, of course, that this is also just the right way to think of people, when we are wondering if they have acted responsibly. But there is a lot to be said for it.

Why do we ask "could he have done otherwise?" We ask it because something has happened that we wish to interpret. An act has been performed, and we wish to understand how the act came about, why it came about, and what meaning we should attach to it. That is, we want to know what conclusions to draw from it about the future. Does it tell us anything about the agent's character, for instance? Does it suggest a criticism of the agent that might, if presented properly, lead the agent to improve his ways in some regard? Can we learn from the incident that this is or is not an agent who can he trusted to behave similarly on similar occasions in the future? If one held his character constant, but changed the circumstances in minor—or even major—ways, would he almost always do the same lamentable sort of thing? Was what we have just observed a "fluke," or was it a manifestation of a "robust" trend—a trend that persists, or is constant, over an interestingly wide variety of conditions?[7]

[6] Shaler Stidham has pointed out to me that in queuing theory, a branch of operations research, there is a method called the common random numbers technique, used on occasion in running simulations of queuing systems to see how well they respond to "random" variation in their operating conditions. In the technique one tests different settings of the design parameters *on the very same sequence of pseudo-random numbers*. This is, in effect, an experimental investigation in "possible worlds," where one test drives slightly different systems in *exactly* the same worlds, and compares their performances. For some purposes this is a provably more sensitive test of relative strengths and weaknesses than the more realistic simulation, in which the different models are test driven on what are "microscopically" different, but "practically" indistinguishable batches of random or pseudo-random numbers. In theoretical work in queuing theory, the assumption of such matched random worlds is known as the assumption of stochastic coupling. For an introductory account, see Fishman 1973.

[7] We are interested in trends and flukes in both directions (praiseworthy and regretted). If we had evidence that Luther was just kidding himself, that his apparently staunch stand was a sort of comic-opera coincidence, our sense of his moral strength would be severely diminished; "He's not so stalwart," we might say, "he could well have done otherwise."

When the agent in question is oneself, this rationale is even more plainly visible. Suppose I find I have done something dreadful. Who cares whether, in exactly the circumstances and state of mind I found myself, I could have done something else? I didn't do something else, and it's too late to undo what I did.[8] But when I go to interpret what I did, what do I learn about myself? Ought I to practice the sort of maneuver I botched, in hopes of it more reliable, less vulnerable to perturbation, or would that be wasted effort? Would it be a good thing, so far as I can tell, for me to try to adjust my habits of thought in such sorts of cases in the future?

Knowing that I will always be somewhat at the mercy of the considerations that merely happen to occur to me as time rushes on, knowing that I cannot entirely control this process of deliberation, I may take steps to bias the likelihood of certain sorts of considerations routinely "coming to mind" in certain critical situations. For instance, I might try to cultivate the habit of counting to ten in my mind before saying anything at all about Ronald Reagan, having learned that the deliberation time thus gained pays off handsomely in cutting down regrettable outbreaks of intemperate commentary. Or I might decide that no matter how engrossed in conversation I am, I must learn to ask myself how many glasses of wine I have had every time I see someone hovering hospitably near my glass with a bottle. This time I made a fool of myself; if the situation had been quite different, I certainly would have done otherwise; if the situation had been virtually the same, I might have done otherwise and I might not. The main thing is to see to it that I will jolly well do otherwise in similar situations in the future.

That, certainly, is the healthy attitude to take toward the regrettable parts of one's recent past. It is the self-applied version of the engineers' attitude toward the persisting weaknesses in the design of the robot. Of course if I would rather find excuses than improve myself, I may dwell on the fact that I don't *have* to "take" responsibility for my action, since I can always imagine a more fine-grained standpoint from which my predicament looms larger than I do. (If you make yourself really small, you can externalize virtually everything.)

In chapter seven I will say more about the rationale for being generous with one's self-ascriptions of responsibility. But for now I will just draw attention to a familiar sort of case in which we hover in the vicinity of asking whether we really could have done otherwise, and then (wisely) back off. One often says, after doing something awful, "I'm terribly sorry; I simply never thought of the consequences; it simply didn't occur to me what harm I was doing!" This looks almost like the beginning of an excuse—"Can I help it what occurs to me and what doesn't?"—but healthy self-controllers shun this path. They *take* responsibility for what might be, very likely is, just an "accident," just one of those things. That way, they make themselves less likely to be "accident" victims in the future.

3. *The Can of Worms*

> The chance of the quantum-theoretician is not the ethical freedom of the Augustinian, and Tyche is as relentless a mistress as Ananke.—Norbert Wiener (1948, P 49)

These edifying reflections invite one final skeptical thrust: "You paint a rosy picture of self-controllers doing the best they can to

[8] I sometimes wonder if part of the subliminal appeal of "radical freedom" or "contra-causal" freedom of choice is this: "I can't change the past (dammit), but I'd feel better about myself if I thought I could almost change it, or, I'd feel better about myself if I learned that it was a sort of cosmic slip for which I was not responsible." It has often been claimed that responsibility and indeterminism are incompatible (for example, Hobart 1934 and Ayer 1954). The argument typically offered is fallacious, as I show in Dennett 1978a, chapter 15. But might it be that this presumed incompatibility of responsibility and "contra-causal freedom" is, secretly, just what attracts some people to contra-causal freedom?

improve their characters, but what sense can be made of this *striving*? If determinism is true, then whatever *does* happen is the only thing that *can* happen." As van Inwagen (1975, pp. 49-50) says, "To deny that men have free will is to assert that what a man does do and what he can do coincide." In a deterministic world what sense could we make of the exhortation to do the best we can? It does seem to us that sometimes people do less well than they are able to do. How can we make sense of this? If determinism is true, and if this means that the only thing one can do is what one does in fact do, then without even trying, everyone will always be doing his very best—and also his very worst. Unless there is some room between the actual and the possible, some elbow room in which to maneuver, we can make no sense of exhortation. Not only that: retrospective judgment and assessment are also apparently rendered pointless. Not only will it be true that everyone always does his best, but everything will be as good as it can be. And as bad. Dr. Pan-gloss, the famous optimist, will be right: it is the best of all possible worlds. But his nemesis, Dr. Pang-loss the pessimist, will sigh and agree: it *is* the best of all possible worlds—and it couldn't be worse![9] As the philosophers' saying goes, "ought" implies "can"—even in domains having nothing whatever to do with free will and moral responsibility.

Even if we are right to abandon allegiance to the "could have done otherwise" principle as a prerequisite of responsible action, there is the residual problem (according to the incompatibilists) that under determinism, we can never do anything but what we in fact do. As Slote observes, "this itself seems a sufficient challenge to deeply entrenched and cherished beliefs to make it worthwhile to see whether the recent arguments can be attacked at some point *before* the conclusion that all actions are necessary." (Slote 1982, p. 9). But the challenge is even more unpalatable than Slote claims. If the incompatibilists were right about us, it would be because they were right about everything: under determinism *nothing* can do anything other than what it in fact does. The conclusion must be that in a deterministic world, since an atom of oxygen that never links up with any hydrogen atoms is determined never to link up with any hydrogen atoms, it is physically *impossible* for it to link up with any hydrogen atoms. In what sense, then, could it be true that it, like any oxygen atom, *can* link up with two hydrogen atoms?

Ayers calls this threatened implication of determinism "actualism"—only the actual is possible. (Ayers 1968, p. 6) Something is surely wrong with actualism, but actualism is so wrong that it is highly unlikely that its falsehood can be parlayed into a *reductio ad absurdum* of determinism. The argument would be disconcertingly short: this oxygen atom has valence 2; therefore it can unite with two hydrogen atoms to form a molecule of water (it *can* right now, whether or not it does); therefore determinism is false. There are impressive arguments from physics that lead to the conclusion that determinism is false—but this isn't one of them.

Hume speaks of "a certain looseness" we want to exist in our world. (*Treatise*, II, III, 2, Selby-Bigge ed., p. 408) This is the looseness that prevents the possible from shrinking tightly around the actual, the looseness presupposed by our use of the word "can." We need this looseness for many things, so we need to know what "can" means, not just for our account of human freedom, and for the social sciences, but for our account of biology, engineering (see chapter three), and in fact any field that relies significantly on statistics and probability theory.

What could the biologist mean, for instance, when speaking of some feature of some species as *better* than some other "possible" feature? If the generally adaptive trend of natural selection is to be coherently described—let alone explained—we must often distinguish a design selected as better (or as *no* better) than other

[9] The wandering "g" in "Pangloss" was pointed out to me by Hofstadter (who also noted that "elbow room" is *almost* "more wobble" backward).

"possible" designs that selection has spurned.[10] Biologists assure us that unicorns are not only not actual; they are impossible—as impossible as mermaids. (It has something to do with the violation of bilaterality required for a single, centered horn, I gather.) But the biologists also assure us that there are many possible species that haven't yet existed, and probably never will—short-legged, fat horses good only for eating, say, or blotchless giraffes. Only a small portion of the possible variations ever appear.

In probability theory, we take it that a coin toss has two possible outcomes: heads or tails.

> When witnessing the toss of a coin, X will normally envisage as possibly true the hypothesis that the coin will land heads up and that it will land tails up. He may also envisage other possibilities—e.g., its landing on its edge. However, if he takes for granted even the crudest folklore of modern physics, he will rule out as impossible the coin's moving upward to outer space in the direction of Alpha Centauri. (Levi 1980, p. 3)

Everywhere one looks, one finds reliance on claims about what things can be in what states, what outcomes are possible, and what is impossible but not *logically* impossible (self-contradictory).

If this elusive sense of "can" has nothing particular to do with agency, it nevertheless makes it appearance vividly in that area. In "Ifs and Cans," Austin (1961) offers a famous series of criticisms of the attempt to define "could have done otherwise" as "would have done otherwise if. . ." for various different fillings of the blank. Austin's objections to this strategy have been ably

[10] See, for instance, Sober's interesting article, "The Evolution of Rationality" (Sober 1981, p. 110), where he speaks of "a selection process in which many possible endowments were simply not represented." Note that the denial of adaptationism, just as much as its assertion, presupposes the coherence of assumptions about possibility. (On the risks and benefits of adaptationism, see Dennett 1983b.)

criticized by several philosophers (see especially Chisholm, 1964a). But more important than those objections and criticisms, which have received a great deal of attention from philosophers, is Austin's abrupt, unargued, and all too influential dismissal (in one footnote and one aside) of the most promising approach to the residual, froggy problem.

Austin notes in passing that "There is some plausibility, for example, in the suggestion that 'I can do X' means 'I shall succeed in doing X, if I try,' and 'I could have done X' means 'I should have succeeded in doing X, if I had tried.'" But a famous long footnote dismisses the suggestion:

> Plausibility, but no more. 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.
>
> But if I tried my hardest, say, and missed, surely there *must* have been *something* that caused me to fail, that made me unable to succeed? So that I *could not* have holed it. Well, a modern belief in science, in there being an explanation of everything, may make us assent to this argument. But such a belief is not in line with the traditional beliefs enshrined in the word *can*: according to *them*, a human ability or power or capacity is inherently liable not to produce success, on occasion, and that for no reason (or are bad luck and bad form sometimes reasons?). (p. 166)

But then what should give way, according to Austin—"a modern belief in science" or the "traditional beliefs enshrined in the word *can*"? Austin does not say, and leaves the impasse unresolved. The impasse is an illusion; modern science needs the same "can" that traditional beliefs about human agency need. And what must give is Austin's insistence that he was "talking about conditions as they precisely were." As we have seen, there is never any purchase to be gained by talking about micro-precise conditions; when we talk about what someone—or something—*can do* we are always interested in something general.

This point is made well by Honoré (1964), in a seldom-cited critical commentary on Austin's paper. Honoré proposes that we distinguish between two senses of "can": "can" (particular) and "can" (general)—and notes that the particular sense is almost degenerate: it "is almost equivalent to 'will' and has predictive force." (p. 464) In the past tense, particular "can" is only appropriate for describing success: "Thus 'I could see you in the undergrowth' is properly said only when I have succeeded in seeing you."

> Success or failure, on the assumption that an effort has been or will be made, is the factor which governs the use of the notion: if the agent tried and failed, he could not do the action: if he tried and succeeded, he was able to do it. (Honoré .1964, p. 464)

The more useful notion is "can" (general), which in the case of an agent imputes skill or ability, and in the case of an inanimate thing, imputes the sort of potentiality discussed in chapter five (for example, the different states that something can be in). But as we saw then, that sense of "can" is a manifestly *epistemic* notion; that is, it is generated by any self-controlling planner's need to partition the world into those things and their "states" that are all possible-for-all-it-knows.

Philosophical tradition distinguishes several varieties of possibility. Among them:

(a) *logical* or "alethic" possibility: the complement of logical impossibility; something is logically possible if it is consistently describable; it is logically possible that there is a unicorn in the garden, but (if the biologists are right) it is not biologically or physically possible.

(b) *physical* or "nomic" possibility: something is physically possible if it does not violate the laws of physics or the laws of nature (*nomos* = law, in Greek). It is physically impossible to travel faster than the speed of light, even though one can describe such a feat without contradicting oneself.

(c) *epistemic* possibility: something is epistemically possible *for Jones* if it is consistent with everything Jones already *knows*. So epistemic possibility is generally viewed as subjective and relative, unlike logical and physical possibility, which are deemed entirely objective.[11]

It is customary in philosophical discussions of free will to distinguish epistemic possibility from its kin, and then dismiss it as of no further interest in that context.[12] Austin's dismissal is one of the briefest. After considering two other senses of "could have," he mentions a third sense,

> in which sense 'I could have done something different' means 'I might, for all anyone could know for certain beforehand, have done something different.' This third kind of 'could have' might, I think, be held to be a vulgarism, 'could' being used incorrectly for 'might': but in any case we shall not be concerned with it here. (Austin 1961, p. 207)

[11] See Hacking 1975 for important complications and qualifications.

[12] See, for example, van Inwagen's brief paragraph (van Inwagen 1983, pp. 9ía), and Ayers' much more cautious approach, which begins: "Discussions of power and potentiality, especially as they occur in the freewill controversy, are fairly haunted by the notion of epistemic possibility, and the related notions of certainty and uncertainty, predictability and unpredictability. This is not mere confusion . . ." (Ayers 1968, pp. 3ff.)

It is a shame that philosophers have not been concerned with it, for it is the key to the resolution of the riddle about "can." The useful notion of "can," the notion that is relied upon not only in personal planning and deliberation, but also in science, is a concept of possibility—and with it, of course, interdefined concepts of impossibility and necessity—that are, contrary to first appearances, fundamentally "epistemic."

As Slote points out in his pioneering article, "Selective Necessity and Free Will" (Slote 1982), the sorts of concepts of necessity and possibility relied upon in these contexts obey different modal principles from the concept of "classical" alethic necessity. In particular, such necessity is not "agglomerative," by which Slote means closed with respect to conjunction introduction.[13] Slote illustrates the concept with an example of an 'accidental' meeting: Jules happens to meet his friend Jim at the bank; he thinks it is a happy accident, as indeed it is. But Jules' being at the bank is not an accident, since he always goes there on Wednesday morning as part of his job; and Jim's being there is also no accident, since he has been sent by his superior. That Jules is at *L* at time *t* is no accident; that Jim is at *L* at time *t* is no accident. But that Jules is at *L* at time *t* and Jim is at *L* at time *t*—that is an accident. (Slote 1982, esp. pp. 15-17)

This is apparently accidentality or coincidentality from-a-limited-point-of-view. We imagine that if we knew much, much more than Jules and Jim together know, we would have been able to predict their convergence at the bank; *to us*, their meeting would have been "no accident." But this is nevertheless just the concept of accidentality we need to describe the "independence" of a thing's powers or abilities from the initial conditions or background conditions in which those powers are exercised. For instance, it is no accident that this particular insect has just the evasive flight pattern it does have (for it was designed by evolution to have that pattern). And it is no accident that the predatory bird that catches that insect has the genes it does (for it too was designed to have those genes). But it is an accident—happy for the bird and its progeny, unhappy for the insect—that a bird with just those genes caught just that evasive insect. And out of thousands of such happy accidents better birds—and better insects—come to be designed. Out of a conspiracy of accidents, by the millions, comes the space of "possibility" within which selection can occur.

The eminent biologist, Jacques Monod, describes the importance for evolution of chance, or what he calls "absolute coincidence" (Monad 1972, p. 12ff.), and illustrates absolute coincidence with an example strikingly like Slote's:

> Suppose that Dr. Brown sets out on an emergency call to a new patient. In the meantime Jones the contractor's man has started making emergency repairs on the roof of a nearby building. As Dr. Brown walks past the building, Jones inadvertently lets go of his hammer, whose (deterministic) trajectory happens to intercept that of the physician, who dies of a fractured skull. We say he was a victim of chance. (p. 114)

But when Monod comes to define the conditions under which such coincidences can occur, he apparently falls into the actualist trap. Accidents must happen if evolution is to take place, Monad says, and accidents can happen—"Unless of course we go back to Laplace's world, from which chance is excluded by definition and where Dr. Brown has been fated to die under Jones' hammer ever since the beginning of time." (p. 115)

[13] Slote overlooks the possibility that the form of "selective necessity" he describes is in fact disguisedly epistemic. But he offers a variety of observations which lead in that direction. Schotch and Jennings (1980) offer several philosophical reasons for doubting the universal appeal of full aggregation or agglomeration principles in modal logic, but miss the most compelling cases, which Slote presents.

If "Laplace's world" means just a deterministic world, then Monod is wrong. Natural selection does not need "absolute" coincidence. It does not need "essential" randomness or perfect independence; it needs practical independence—of the sort exhibited by Brown and Jones, and Jules and Jim, each on his own trajectory but "just happening" to intersect, like the cards being deterministically shuffled in a deck and just happening to fall into sequence. Would evolution occur in a deterministic world, a Laplacean world where mutation was caused by a nonrandom process? Yes, for what evolution requires is an unpatterned generator of raw material, not an uncaused generator of raw material. Quantum-level effects may indeed play a role in the generation of mutations, but such a role is not required by theory.[14]

It is not clear that "genuine" or "objective" randomness of either the quantum-mechanical sort or of the mathematical, informationally incompressible sort is ever required by a process, or detectable by a process. (Chaitin (1976) presents a Gödelian proof that there is no decision procedure for determining whether a series is mathematically random.) Even in mathematics, where the concept of objective randomness can find critical application within proofs, there are cases of what might be called practical indistinguishability.

In number theory, the Fermat-Lagrange Theorem states that every natural number is the sum of four perfect squares:

$$n^2 = x^2 + y^2 + z^2 + w^2$$

The theorem is easy enough to prove, I gather, but finding the values for *x*, *y*, *z*, and *w* for a given n is a tedious business. There is a straightforward, "brute force" algorithm that will always find the values by simple exhaustive trial and error, but it has the alarming property of requiring, on the average, 2^n steps to terminate. Thus, for a natural number as small as, say, 203, the algorithm could not be expected to find the answer before the heat death of the universe. It is not what the jargon calls a *feasible algorithm*, even though *in principle* (as a philosopher would note) it always yields the correct answer.

But all is not lost. Rabin and others have developed so-called random algorithms, which rely in extremely counterintuitive ways on randomization. One such algorithm has been discovered by Rabin for finding values for the Fermat-Lagrange theorem. It is not logically guaranteed to find the right answer any faster than the brute force algorithm, but its *expected* termination time (with the right answer) is only $(\log n)^3$ steps, a manageably small number even for large values of *n*. The probability of a much longer or much shorter termination time drops off so steeply as to be entirely negligible. The formal proof that this is its expected termination time makes essential mention of the invocation of random sequences in the algorithm.

Question: in the actual world of hardware computers, does it make any difference whether the computer uses a genuinely random sequence or a pseudo-random sequence? That is, if one wrote Rabin's program to run on a computer that didn't have a radium randomizer but relied instead on a pseudo-random number generating algorithm, would this cheap shortcut work? Or would attempts to find the values for a particular *n* run longer than the expected number of steps in virtue of the hidden, humanly undetectable nonrandomness of the sequence? Would the number system, in its hauteur, punish the mathematician for trying to plumb its secrets with mere pseudo-random exploration? As it turns out, experience to date has been that one can indeed get away

[14] Monod's very interesting discussion (see esp. pp. 77-80 and 111-117) is equivocal and conflicted on this point. See also the valuable discussion of this question in Wimsatt 1980, esp. section 3, "Periodic, Almost-Periodic, and Chaotic Behavior in Simple Models of Population Growth and Regulation."

with pseudorandom sequences. In the actual runs that have been attempted, it has made no difference.[15]

But surely mere practical indistinguishability, even in the limit, is not the Real Thing—real, objective possibility. That is the intuition we must now examine. It is at the heart of the brusque rejection, by philosophers, of epistemic possibility as a building stone in the foundation of free will. So-called "classical" or Newtonian physics is deterministic, but as several physicists have recently noted, many of the most mundane macroscopic phenomena in a Newtonian world would be, *by Newtonian principles*, unpredictable by any being that fell short of being an *infinite* Laplacean demon, for they would require infinite precision of initial observation. That is, errors in observation, however minuscule, would propagate and grow exponentially (Berry 1983 and Ford 1983).

In Newtonian physics, there are stable systems (precious few of them) and unstable or chaotic systems. "For nonchaotic systems, error propagates less rapidly and . . . even a coarse-grained past suffices to determine precisely a coarse-grained future." Eclipses, for instance, may be predicted centuries in advance. But "a chaotic orbit is random and incalculable; its information content is both infinite and incompressible." (Ford 1983, p. 7) The trajectory of a pinball (the example is Berry's) after bumping, say, twenty posts (in a few seconds) is unpredictable *in the limit*, far outstripping the limits of accuracy of any imaginable observation devices. Now this result is surely "just epistemic." What could it have to do with free will?

Just this, I think: such chaotic systems are the source of the "practical" (but one might say infinitely practical) independence of things that shuffles the world and makes it a place of continual opportunity. The opportunities provided are not just *our* opportunities, but also those of Mother Nature—and of oxygen atoms which can join forces on occasion with hydrogen atoms. It is not any parochial fact about our epistemic limitations that distinguishes the world into stable, predictable systems and unstable, chaotic systems; it is a fact about the world itself—because it is a fact about the world's predictability by any predicting system at all, however powerful. There is no higher perspective (unless we count the perspective of an infinite being) from which the "accidental" collisions of locally predictable trajectories are themselves predictable and hence "no accident" after all.

It is this contrast between the stable and the chaotic that grounds our division of the world into the enduring and salient features of the world, and those features that we *must* treat statistically or probabilistically (in effect, either averaging over them and turning them into a blur, or treating them as equi-possible members of some ensemble of alternatives). And this division of the world is not just our division; it is, for instance, Mother Nature's division as well. Since for all Mother Nature knows (or could know) it is possible that these insects will cross paths (sometime, somewhere) with these insectivorous birds, they had better be designed with some avoidance machinery. This endows them with a certain power (a bit of "can do," as slang has it) that will serve well (in general).

(These all too sketchy remarks about "can" are at best a pointing gesture toward the final, finished surface of this part of my sculpted portrayal of the free agent. This is another area where much more work needs to be done, and some of the work, certainly, is quite beyond me. But if I am even approximately right in this first, rough pass over the region, the work still to be done will at least move the investigation off of stale, overworked surfaces into new spaces.)

[15] Rabin, personal communication, New York Academy of Sciences meeting, April 1983. See also Rabin 1980 and Jauch 1973 (discussed in Hofstadter 1979, pp. 408-409).