3
Responsiveness and Moral Responsibility

We distinguish between creatures who can legitimately be held morally responsible for their actions and those who cannot. Among the actions a morally responsible agent performs, we distinguish between those actions for which the agent is morally responsible and those for which he is not.

An agent is morally responsible for an action insofar as he is rationally accessible to certain kinds of attitudes and activities as a result of performing the action. The attitudes include resentment, indignation, respect, and gratitude; and the activities include moral praise and blame, and reward and punishment.[1] With this approach, an agent can be a rational candidate for praise or blame, even though he is neither praiseworthy nor blameworthy. For instance, an agent can be morally responsible for a morally "neutral" act. A theory of moral responsibility sets the conditions under which we believe that an individual is a *rational candidate* for praise or blame on account of his behavior. This theory needs to be supplemented by a further moral theory that specifies which agents, among those who are morally responsible, *ought* to be praised or blamed (and to what extent) for their actions. Whereas both kinds of theory are obviously important, I focus here on the first sort of theory—one that explains rational accessibility to the pertinent attitudes and activities.

What I present here is really just a sketch of a theory. It needs to be elaborated and defended much more carefully and explicitly. But I hope that enough of its content will be presented to see that it is a worthwhile approach to develop. The kind of theory I present is certainly not radically new and entirely different from its predecessors.[2] But I hope to develop the theory in a way that avoids some of the objections to similar approaches, and I will draw out some implications that have so far gone unnoticed.

I have benefited greatly from comments on previous versions of this paper by Sarah Buss, Anthony Brueckner, and Ferdinand Schoeman. I also benefited from reading a version of this paper at Birkbeck College, University of London.

A Sketch of a Theory of Moral Responsibility

A theory of moral responsibility should capture our intuitive judgments about clear cases. That is, I assume there is at least fairly wide agreement about certain cases in which an agent can reasonably be held morally responsible for what he does and certain cases in which an agent cannot be held responsible. Considered opinions about these sorts of situations are important data to be explained by a theory of moral responsibility. In order to generate a principle that might underlie our reactions to relatively clear cases, it is useful to begin by considering examples in which we are inclined to think that an agent cannot legitimately be held morally responsible.

Imagine that an individual has been hypnotized. The hypnotist has induced an urge to punch the nearest person after hearing the telephone ring. Insofar as the individual did not consent to this sort of hypnotic suggestion (perhaps he has undergone hypnosis to help him stop smoking), it seems unreasonable to hold him morally responsible for punching his friend in the nose upon hearing the telephone ring.

Suppose similarly that an evil person has got hold of Smith's television set and has wired it so as to allow him to subject Smith to a sophisticated sort of subliminal advertising. The bad person systematically subjects Smith to subliminal advertising that causes Smith to murder his neighbor. Because of the nature of the causal history of the action, it is apparent that Smith cannot be held morally responsible for the lamentable deed.

We feel similarly about actions produced in a wide variety of ways. Agents who perform actions produced by powerful forms of brainwashing and indoctrination, potent drugs, and certain sorts of direct manipulation of the brain are not reasonably to be held morally responsible for their actions. Imagine, for instance, that neurophysiologists of the future can isolate certain key parts of the brain, which can be manipulated in order to induce decisions and actions. If scientists electronically stimulate those parts of Jones's brain, thus causing him to help a person who is being mugged, Jones himself cannot reasonably be held morally responsible for his behavior. It is not to Jones's credit that he has prevented a mugging.

Also, if we discover that a piece of behavior is attributable to a significant brain lesion or a neurological disorder, we do not hold the agent morally responsible for it. Similarly, certain sorts of mental disorders—extreme phobias, for instance—may issue in behavior for which the agent cannot reasonably be held responsible.

Many people feel there can be genuinely "irresistible" psychological impulses. If so, then these may result in behavior for which the agent cannot be held morally responsible. Drug addicts may (in certain circumstances) act on literally irresistible urges, and we might not hold them morally responsible for acting on these desires (especially if we believe they are not morally responsible for acquiring the addiction in the first place).

Also, certain sorts of coercive threats (and perhaps offers) rule out moral responsibility. The bank teller who is told he will be shot unless he hands over the money may have an overwhelming and irresistible desire to comply with the

demand. Insofar as he acts from such an impulse, it is plausible to suppose that the teller is not morally responsible for his action.[3]

Evidently, the causal history of an action matters to us in making moral responsibility attributions. When persons are manipulated in certain ways, they are like marionettes and are not appropriate candidates for praise or blame. Certain factors issuing in behavior are, we understand intuitively, responsibility-undermining factors.

We can contrast such cases—in which some responsibility-undermining factor operates—with cases in which there is the "normal," unimpaired operation of the human deliberative mechanism. When you deliberate about whether to give 5 percent of your salary to the United Way and consider reasons on both sides, and your decision to give the money is not induced by hypnosis, brainwashing, direct manipulation, psychotic impulses, and so on, we think you can legitimately be praised for your charitable action. Insofar as we can identify no responsibility-undermining factor at work in your decision and action, we are inclined to hold you morally responsible.

Now it might be thought that there is a fairly obvious way of distinguishing the clear cases of moral responsibility from the clear cases of lack of it. It seems that, in the cases in which an agent is morally responsible for an action, he is free to do otherwise, and in the cases of lack of moral responsibility, the agent is not free to do otherwise. Thus, it appears that the actual operation of what is intuitively a responsibility-undermining factor rules out moral responsibility because it rules out freedom to do otherwise.

The point could be put as follows. When an agent is (for example) hypnotized, he is not sensitive to reasons in the appropriate way. Given the hypnosis, he would still behave in the same way no matter what the relevant reasons were. Suppose, again, that an individual is hypnotically induced to punch the nearest person after hearing the telephone ring. Now given this sort of hypnosis, he would punch the nearest person after hearing the telephone ring, even if he had extremely strong reasons not to. The agent here is not responsive to reasons—the behavior would be the same no matter what reasons there were.

In contrast, when there is the normal, unimpaired operation of the human deliberative mechanism, we suppose that the agent *is* responsive to reasons. So when you decide to give money to the United Way, we think that you nevertheless would not have contributed had you discovered that there was widespread fraud within the agency. Thus it is very natural and reasonable to think that the difference between morally responsible agents and those who are not consists in the "reasons-responsiveness" of the agents.

But I believe that there are cases in which an agent can be held morally responsible for performing an action, even though that person could not have done otherwise (and is not "reasons-responsive").[4] Here is a graphic example. Imagine that an evil person has installed a device in Brown's brain which allows him to monitor Brown's mental activity and also to intervene in it, if he wishes. He can electronically manipulate Brown's brain by "remote control" to induce decisions, and let us imagine that he can also ensure that Brown acts on the decisions so induced. Now suppose that Brown is about to murder his neighbor, and that this is precisely what the evil person wishes. That is, let us imagine that the device simply

monitors Brown's brain activity, but that it plays no role in Brown's actual decision and action. Brown deliberates and behaves just as he would have if no device had been implanted in his brain. But we also imagine that had Brown begun to decide not to murder his neighbor, the device would have been activated and would have caused him to choose to murder the neighbor (and to do so) anyway. Here is a case where an agent can be held morally responsible for performing an action, although he could not have done otherwise.[5] Let us call such a case a "Frankfurt-type" case.

In a Frankfurt-type case, the actual sequence proceeds in a way that grounds moral responsibility attributions, even though the alternative scenario (or perhaps a range of alternative scenarios) proceeds in a way that rules out responsibility. In a Frankfurt-type case, no responsibility-undermining factor occurs in the actual sequence, although such a factor occurs in the alternative scenario. Such cases impel us to adopt a more refined theory of moral responsibility—an "actual-sequence model" of moral responsibility. With such an approach, we distinguish between the kinds of mechanisms that operate in the actual sequence and in the alternative sequence (or sequences).

In a Frankfurt-type case, the kind of mechanism that actually operates is reasons-responsive, although the kind of mechanism that would operate in the alternative scenario is *not*.[6] In the case discussed above, Brown's action issues from the normal faculty of practical reasoning, which we can reasonably take to be reasons-responsive. But in the alternative scenario, a different kind of mechanism would have operated—one involving direct electronic stimulation of Brown's brain. And this mechanism is not reasons-responsive. Thus, the actual-sequence mechanism can be reasons-responsive, even though the *agent* is not reasons-responsive. (*Brown* could not have done otherwise.)

The suggestion, then, for a more refined way of distinguishing the relatively clear cases of moral responsibility from cases of the lack of it is as follows. An agent is morally responsible for performing an action insofar as the mechanism that actually issues in the action is reasons-responsive. When an unresponsive mechanism actually operates, it is true that the agent is not free to do otherwise; but an agent who is unable to do otherwise may act from a responsive mechanism and can thus be held morally responsible for what he does.

So far I have pointed to some cases in which it is intuitively clear that a person cannot be held morally responsible for what he has done and other cases in which it is intuitively clear that an agent can be held responsible. I have suggested a principle that might distinguish the two types of cases. This principle makes use of two ingredients: reasons-responsiveness and the distinction between actual-sequence and alternative-sequence mechanisms. But I have been somewhat vague and breezy about formulating the principle. It is now necessary to explain it more carefully, beginning with the notion of reasons-responsiveness.

Reasons-Responsiveness

I wish to discuss two kinds of reasons-responsiveness: strong and weak. Let's begin with strong reasons-responsiveness. Strong reasons-responsiveness obtains when a

certain kind *K* of mechanism actually issues in an action and if there were sufficient reason to do otherwise and *K* were to operate, the agent would recognize the sufficient reason to do otherwise and thus choose to do otherwise and do otherwise. To test whether a kind of mechanism is strongly reasons-responsive, one asks what would happen if there were sufficient reason for the agent to do otherwise and the actual-sequence mechanism were to operate. Under circumstances in which there are sufficient reasons for the agent to do otherwise and the actual type of mechanism operates, three conditions must be satisfied: The agent must take the reasons to be sufficient, choose in accordance with the sufficient reason, and act in accordance with the choice. Thus, there can be at least three sorts of "alternative-sequence" failures: failures in the connection between what reasons there are and what reasons the agent recognizes, in the connection between the agent's reason and choice, and in the connection between choice and action.

The first kind of failure is a failure to be *receptive* to reasons. It is the kind of inability that afflicts certain delusional psychotics.[7] The second kind of failure is a failure of *reactivity*—a failure to be appropriately affected by beliefs. Lack of reactivity afflicts certain compulsive or phobic neurotics.[8] Finally, there is the failure successfully to translate one's choice into action; this failure is a kind of impotence. If none of these failures were to occur in the alternative sequence (and the actual kind of mechanism were to operate), then the actually operative mechanism would be strongly reasons-responsive. There would be a tight fit between the reasons there are and the reasons the agent has, the agent's reasons and choice, and choice and action. The agent's actions would fit the contours of reasons *closely*.[9]

I believe that, when an action issues from a strongly reasons-responsive mechanism, this suffices for moral responsibility; but I do not believe that strong reasons-responsiveness is a necessary condition for moral responsibility. To see this, imagine that as a result of the unimpaired operation of the normal human faculty of practical reasoning, I decide to go (and go) to the basketball game tonight, and that I have sufficient reason to do so; but suppose that I would have been "weak-willed" had there been sufficient reason *not* to go. That is, imagine that had there been a sufficient reason not to go, it would have been that I had a strict deadline for an important manuscript (which I could not meet, if I were to go to the game). I nevertheless would have chosen to go to the game, even though I would have recognized that I had sufficient reason to stay home and work. It seems to me that I actually go to the basketball game freely and can reasonably be held morally responsible for going; and yet the actual-sequence mechanism that results in my action is not reasons-responsive in the strong sense. The failure of strong reasons-responsiveness here stems from my disposition toward weakness of the will.

Going to the basketball game is plausibly thought to be a morally neutral act; in the approach to moral responsibility adopted here, one can be morally responsible for an action, even though the act is neither praiseworthy nor blameworthy. The phenomenon of weakness of will also poses a problem for intuitively clear cases of moral responsibility for *commendable* acts. Suppose, for example, that I devote my afternoon to working for the United Way (and my decision and action proceed via an intuitively responsibility-conferring mechanism). And imagine

that, if I had a sufficient reason to refrain, it would (again) have been my publication deadline. But imagine that I would have devoted my time to charity even if I had such a reason not to. Here it seems that I am both morally responsible and praiseworthy for doing what I do, and yet the actual mechanism is not strongly reasons-responsive.

Further, it is quite clear that strong reasons-responsiveness cannot be a necessary condition for moral responsibility for morally blameworthy and/or imprudent acts. Suppose that I steal a book from a store, knowing full well that it is morally wrong for me to do so and that I will be apprehended and thus that it is not prudent of me to do so. Nevertheless, the actual sequence may be intuitively responsibility-conferring; no factors that intuitively undermine moral responsibility may actually operate. (Of course, I assume that there can be genuine cases of weak-willed actions that are free actions for which the agent can be held responsible.) Here, then, is a case in which I am morally responsible for stealing the book, but my actual-sequence mechanism is not strongly reasons-responsive: There actually is sufficient reason (both moral and prudential) to do otherwise, and yet I steal the book.

All three cases presented above provide problems for the claim that strong reasons-responsiveness is necessary for moral responsibility. Strong reasons-responsiveness may be both sufficient and necessary for a certain kind of praiseworthiness—it is a great virtue to connect one's actions with the contours of value in a strongly reasons-responsive way. Of course, not all agents who are morally responsible are morally commendable (or even maximally prudent). I believe that moral responsibility requires only a looser kind of fit between reasons and action: "weak reasons-responsiveness."

Under the requirement of strong reasons-responsiveness, we ask what would happen if there were a sufficient reason to do otherwise (holding fixed the actual kind of mechanism). Strong reasons responsiveness points us to the alternative scenario in which there is a sufficient reason for the agent to do otherwise (and the actual mechanism operates), which is *most similar* to the actual situation. Put in terms of possible worlds, the nonactual possible worlds that are germane to strong reasons-responsiveness are those in which the agent has a sufficient reason to do otherwise (and in which the actual kind of mechanism operates) that are most similar to the actual world. (Perhaps there is just one such world, or perhaps there is a sphere of many such worlds.) In contrast, under weak reasons-responsiveness, there must exist *some* possible world in which there is a sufficient reason to do otherwise, the agent's actual mechanism operates, and the agent does otherwise. This possible world need not be the one (or ones) in which the agent has a sufficient reason to do otherwise (and the actual mechanism operates), which is (or are) *most similar* to the actual world.[10]

Consider again my decision to go to the basketball game. In this situation, if I were to have a sufficient reason to do otherwise, this would be a publication deadline; and I would under such circumstances be weak-willed and still go to the game. However, there certainly exists *some* scenario in which the actual mechanism operates, I have sufficient reason not to go to the game, and I don't go. Suppose, for instance, that I am told that I will have to pay $1,000 for a ticket to the

game. Even though I am disposed to be weak-willed under some circumstances, there are some circumstances in which I would respond appropriately to sufficient reasons. These are circumstances in which the reasons are considerably *stronger* than the reasons which would exist if I were to have sufficient reason to do otherwise.

Consider, similarly, my commendable act of working this afternoon for the United Way. Even though I would do so anyway, even if I had a publication deadline, I certainly would *not* work for the United Way if to do so I would have to sacrifice my job. Thus, the actual mechanism issuing in my action is weakly reasons-responsive. Also, when an agent wrongly (and imprudently) steals a book (i.e., there actually is sufficient reason not to), the actual mechanism might be responsive to at least some logically possible incentive not to steal. To the extent that it is so responsive, he is properly held morally responsible for stealing the book. Even an agent who acts against good reasons can be responsive to *some* reasons.

I believe that the agent's actual-sequence mechanism *must* be weakly reasons-responsive if he is to be held morally responsible. If (given the operation of the actual kind of mechanism) he would persist in stealing the book even knowing that by so acting he would cause himself and his family to be killed, then the actual mechanism would seem to be inconsistent with holding that person morally responsible for an action.

An agent whose act is produced by a strongly reasons-responsive mechanism is commendable; his behavior fits tightly the contours of value. But a weakly responsive mechanism is all that is required for moral responsibility. In my approach, actual irrationality is compatible with moral responsibility (as it should be). Perhaps Dostoyevsky's underground man is an example of an actually irrational and yet morally responsible individual. Similarly, certain kinds of hypothetical irrationality are compatible with moral responsibility; a tendency toward weakness of the will need not point to any defect in the actual mechanism leading to action. Moral responsibility requires *some* connection between reason and action, but the fit can be quite loose.[11]

In this section I have distinguished two kinds of responsiveness. I have argued that an agent is morally responsible for an action insofar as the action is produced by a weakly reasons-responsive mechanism. In the next section, I discuss an analogy between this theory of moral responsibility and a parallel sort of theory of knowledge. This analogy will help to refine our understanding of the actual-sequence nature of moral responsibility. In the following section, I further sharpen the formulation of the theory by rendering more precise the key idea of a "kind of mechanism issuing in action."

Knowledge and Responsibility

I have sketched an actual-sequence model of moral responsibility. In this approach, an agent can be morally responsible for performing an action although he is not free to do otherwise. It is sufficient that the actual-sequence mechanism be responsive to reasons in the appropriate way. There is an analogy between this sort of theory of moral responsibility and an actual-sequence model of knowledge. In

this approach to knowledge, an agent may have knowledge of a certain proposition, even though he lacks the pertinent discriminatory capacity. It is sufficient that the actual-sequence mechanism be sensitive to truth in the appropriate way.

In order for a person to know that *p*, it is clear that the person must believe that *p*, and that *p* must be true; but this is surely not enough, and there are various strategies for providing further requirements.[12] One "externalist" approach claims that the person's belief that *p* must be a "reliable indicator" of *p*'s truth—or perhaps, that it must "track" *p*'s truth. Very roughly, one might say that, in order for an agent to have knowledge that *p*, it must be the case both that (1) the agent would not believe that *p* if *p* were not true, and (2) under various conditions in which *p* were true, the agent would believe that *p*. One asks here about the agent's beliefs in a sphere of worlds that are relatively similar to the actual world—both worlds in which *p* is true and worlds in which *p* is false.[13]

So suppose that as you are driving along, you see what you take to be a barn in a field, and that you conclude that it is a barn in the field; and it is an ordinary barn in a field. Unknown to you, had it not been a barn, a demonic farmer would have installed a papier-mâché replica of a barn. In this case you truly believe that it is a normal barn in the field, but your belief does not "track truth": had there been no barn in the field, you still would have believed there to be a barn in the field. In this case you lack a discriminatory capacity that might seem required for knowledge.

Let us contrast this case with another in which you see a banana in a supermarket, and you conclude that there is a banana on the shelf. We suppose here that there is no demonic supermarket manager poised to fool you, and that if there were no banana on the shelf, you would not believe that there is a banana on the shelf. Presumably, in this case your belief tracks truth, and you might be said to know that there is a banana on the shelf. Furthermore, this is so even though *there exists* a logically possible scenario in which a demonic supermarket manager has placed a plastic banana on the shelf and you still conclude that it is a banana. In this account, what is pertinent to knowledge are the scenarios in which *p* is false that are *most similar* to the actual world; that there are more remote possibilities in which the proposition *p* is false is not taken by the approach to be germane to whether the individual has knowledge.[14]

The cases described above might suggest that an agent has knowledge that *p* only if he has the ability to discriminate the conditions that would obtain if *p* were true from those that would obtain if *p* were false. However, consider the following examples (from Nozick):

> A grandmother sees her grandson is well when he comes to visit; but if he were sick or dead, others would tell her he was well to spare her upset. Yet this does not mean she doesn't know he is well (or at least ambulatory) when she sees him.[15]
>
> S believes a certain building is a theater and concert hall. He has attended plays and concerts there. . . . However, if the building were not a theater, it would have housed a nuclear reactor that would so have altered the air around it (let us suppose) that everyone upon approaching the theater would have become lethargic and nauseous, and given up the attempt to buy a ticket. The government cover story would

> have been that the building was a theater, a cover story they knew would be safe since no unmedicated person could approach through the nausea field to discover any differently. Everyone, let us suppose, would have believed the cover story; they would have believed that the building they saw (but only from some distance) was a theater.[16]

These examples are epistemological analogues to Frankfurt-type cases in which an agent is morally responsible for performing an action although he could not have done otherwise. In these cases an agent knows that *p*, although he lacks the pertinent discriminatory capacity. Just as we switched from demanding agent-responsiveness to demanding mechanism-responsiveness for moral responsibility, it is appropriate to demand only mechanism-sensitivity to truth in order for an agent to have knowledge.

As Nozick points out, it is possible to believe that *p* via a truth-sensitive mechanism, and thus know that *p*, even though an insensitive mechanism would have operated in the alternative scenario (or scenarios). Thus, we want an actual-sequence theory of knowledge, just as we want an actual-sequence theory of responsibility. We need to distinguish between actual-sequence and alternative-sequence mechanisms and focus on the properties of the actual-sequence mechanism. But whereas there is a strong analogy between the theories of responsibility and knowledge sketched above, I now want to point to two important differences between responsibility and knowledge.

First, in the theory of responsibility presented above, if an agent acts on a mechanism of type M, there must be *some* possible scenario in which M operates, the agent has sufficient reason to do otherwise, and he does do otherwise, in order for the agent to be morally responsible for his action. The possible scenario need not be the one that would have occurred if M had operated and the agent had sufficient reason to do otherwise. That is, the scenario pertinent to responsibility ascriptions need not be the scenario (or set of them) in which an M-type mechanism operates and the agent has sufficient reason to do otherwise that are *most similar* to the actual scenario. In contrast, in the theory of knowledge presented above, if an agent believes that *p* via an M-type mechanism, then it must be the case that if an M-type mechanism were to operate and *p* were false, the agent would believe that *p* is false if the agent is to know that *p*.

Roughly speaking, the logical possibilities pertinent to moral responsibility attributions may be more remote than those pertinent to knowledge attributions. I believe, then, that the connection between reasons and action that is necessary for moral responsibility is "looser" than the connection between truth and belief that is necessary for knowledge. Of course, this point is consistent with the claim that both knowledge and moral responsibility are actual-sequence notions; it is just that actual-sequence truth-sensitivity is defined more "strictly" (i.e., in terms of "closer" possibilities) than actual-sequence reasons-responsiveness.

But I believe there is a second difference between moral responsibility and knowledge. I have claimed that, just as moral responsibility does not require freedom to do otherwise, knowledge does not require the capacity to discriminate; what is sufficient in the case of responsibility is reasons-responsiveness, and in the

case of knowledge, truth-sensitivity. Thus both notions are actual-sequence notions. But I wish to point out a stronger sense in which moral responsibility (and not knowledge) depends only on the actual sequence.

I claim that an agent's moral responsibility for an action is supervenient on the actual physical causal influences that issue in the action, whereas an agent's knowledge that *p* is *not* supervenient on the actual physical causal influences that issue in the belief that *p*. First, let me explain the supervenience claim for moral responsibility. It seems to me impossible that there be cases in which there are two agents who perform actions of the same type as a result of exactly the same kind of actual causal sequence, but in which one agent is morally responsible for the action and the other is *not*. Differences in responsibility ascriptions must come from differences in the actual physical factors resulting in action; mere differences in alternate scenarios do not translate into differences in responsibility ascriptions. That is, differences in responsibility ascriptions must come from differences in the actual histories of actions, and not mere "possible" histories.

Suppose you and I both heroically jump into the lake to save a drowning swimmer, and everything that actually happens in both cases is relevantly similar—except that whereas you could have done otherwise, I could not have. (I could not have done otherwise by virtue of the existence of a mechanism in my brain that would have stimulated it to produce a decision to save the swimmer had I been inclined not to.) Insofar as the actual physical sequences issuing in our behavior are the same, we are equally morally responsible.

However, here is an epistemological example of Nozick's:

> Consider another case, of a student who, when his philosophy class is cancelled, usually returns to his room and takes hallucinogenic drugs; one hallucination he has sometimes is of being in his philosophy class. When the student actually is in the philosophy class, does he know he is? I think not, for if he weren't in class, he still might believe he was. . . . Two students in the class might be in the same actual situation, having (roughly) the same retinal and aural intake, yet the first knows he is in class while the other does not, because they are situated differently subjunctively—different subjunctives hold true of them.[17]

The two students have exactly the same actual physical factors issue in beliefs that they are in class. However, one student does not know he is in class: if he were not in class (and he were to employ the method of introspection, which was actually employed), then he would (or at least might) still believe that he is in class (as a result of the drug). The other student—who is not disposed to use the drug—does know that he is in class. Thus knowledge is not supervenient on actual physical facts in the way that moral responsibility is.

I have claimed above that there is a certain parallel between moral responsibility and knowledge: The reasons-responsiveness of the actual mechanism leading to action suffices for responsibility, and the truth sensitivity of the actual mechanism leading to belief suffices for knowledge. How exactly is this claim of parallelism compatible with the further claim that moral responsibility attributions are supervenient on actual physical causal factors, whereas knowledge attributions are *not*? I think the answer lies in our intuitive way of individuating "mechanisms."

We tend to individuate mechanisms more finely in action theory than in epistemology.

In the case of the first student, we take the relevant mechanism issuing in belief to be "introspection." Of course, the same sort of mechanism would have operated had the student taken the drug. With this "wide" kind of individuation of mechanisms, it turns out that the mechanism that issues in the one student's belief is *not* truth-sensitive, whereas the mechanism of the other student *is*.

However, in the case in which I save the drowning child ("on my own"), it is natural to suppose that if I had been stimulated by the scientists, this would have been a kind of mechanism *different* from the one that actually operates. Similarly, had I been injected with a drug that issued in an irresistible desire to save the drowning swimmer, this would have constituted a kind of mechanism *different* from the actual one. With this "narrow" kind of individuation of mechanisms, it turns out that the mechanism that issues in my action of saving the child *is* reasons-responsive (just as yours is).

The asymmetry of supervenience is compatible with the symmetrically actual-sequence nature of knowledge and moral responsibility. The asymmetry of supervenience is generated by the intuitively natural tendency to individuate mechanisms issuing in belief more broadly than mechanisms issuing in action.[18]

Mechanisms

I have suggested that an agent is morally responsible for performing an action insofar as the mechanism that actually issues in the action is reasons-responsive; but this suggestion needs to be refined in light of the fact that various different mechanisms may actually operate in a given case. Which mechanism is relevant to responsibility ascriptions?

Suppose that I deliberate (in the normal way) about whether to donate 5 percent of my paycheck to the United Way, and that I decide to make the donation and act on my decision. We might fill in the story so that it is intuitively a paradigmatic case in which I am morally responsible for my action; and yet consider the actually operative mechanism, "deliberation preceding donating 5 percent of one's salary to the United Way." If *this* kind of mechanism were to operate, then I would give 5 percent of my paycheck to the United Way in any logically possible scenario. Thus, this kind of actually operative mechanism is *not* reasons-responsive.

However, a mechanism such as "deliberating prior to giving 5 percent of one's salary to the United Way" is not of the kind that is relevant to moral responsibility ascriptions. This is because it is not a "temporally intrinsic" mechanism. The operation of a temporally extrinsic or "relational" mechanism already includes the occurrence of the action it is supposed to cause.

Note that the operation of a mechanism of the kind "deliberating prior to giving 5 percent of one's paycheck to the United Way" *entails* that one give 5 percent of one's paycheck to the United Way. In this sense, then, the mechanism already includes the action: its operation entails that the action occurs. Thus, it is a necessary condition of a mechanism's relevance to moral responsibility ascriptions

(on the theory proposed here) that it be a "temporally intrinsic" or "nonrelational" mechanism in the following sense: if a mechanism M issues in act X, then M is relevant to the agent's moral responsibility for performing X only if M's operating does not entail that X occurs. I believe that the requirement that a mechanism be temporally intrinsic is an intuitively natural and unobjectionable one. Of course, we have so far only a necessary condition for being a relevant mechanism; there may be various different mechanisms that issue in an action, all of which are temporally intrinsic. Which mechanism is "the" mechanism pertinent to moral responsibility ascription?

I do not have a theory that will specify in a general way how to determine which mechanism is "the" mechanism relevant to assessment of responsibility. It is simply a presupposition of this theory as presented above that, for each act, an intuitively natural mechanism is appropriately selected as *the* mechanism that issues in action, for the purposes of assessing moral responsibility.

I do not think this presupposition is problematic. But if there is a worry, it is useful to note that the basic theory can be formulated without such a presupposition. As so far developed, the theory says that an agent is morally responsible for performing an action insofar as the (relevant, temporally intrinsic) mechanism issuing in the action is reasons-responsive. Alternatively, one could say that an agent is morally responsible for an action insofar as there is no actually operative temporally intrinsic mechanism issuing in the action that is not reasons-responsive. This alternative formulation obviates the need to select one mechanism as the "relevant" one. In what follows I continue to employ the first formulation, but the basic points should apply equally to the alternative formulation.

I wish now to apply the theory to a few cases. We think intuitively that irresistible urges can be psychologically compulsive and can rule out moral responsibility. Imagine that Jim has a literally irresistible urge to take a certain drug, and that he does in fact take the drug. What exactly is the relevant mechanism that issues in Jim's taking the drug? Notice that the mechanism "deliberation involving an irresistible urge to take the drug" is not temporally intrinsic and thus not admissible as a mechanism pertinent to moral responsibility ascription: its operation entails that Jim takes the drug. Consider, then, the mechanism "deliberation involving an irresistible desire." Whereas this mechanism *is* temporally intrinsic, it is also reasons-responsive: There is a possible scenario in which Jim acts on this kind of mechanism and refrains from taking the drug. In this scenario, Jim has an irresistible urge to *refrain* from taking the drug. These considerations show that neither "deliberation involving an irresistible desire for the drug" nor "deliberation involving an irresistible desire" is the relevant mechanism (if the theory of responsibility is to achieve an adequate fit with our intuitive judgments).

When Jim acts on an irresistible urge to take the drug, there is some physical process of kind *P* taking place in his central nervous system. When a person undergoes this kind of physical process, we say that the urge is literally irresistible. I believe that what underlies our intuitive claim that Jim is not morally responsible for taking the drug is that the relevant kind of mechanism issuing in Jim's taking the drug is of physical kind *P*, and that a mechanism of kind *P* is not reasons-responsive. When an agent acts from a literally irresistible urge, he is undergoing a kind of physical process that is not reasons-responsive, and it is this lack of

reasons-responsiveness of the actual physical process that rules out moral responsibility.[19]

Consider again my claim that certain sorts of "direct manipulation of the brain" rule out moral responsibility. It is clear that not all such manipulations would rule out moral responsibility. Suppose, for instance, that a scientist manipulates just one brain cell at the periphery of my brain. This kind of manipulation need not rule out responsibility insofar as this kind of physical process can be reasons-responsive. It is when the scientists intervene and manipulate the brain in a way which is *not* reasons-responsive that they undermine an agent's moral responsibility for action.[20]

Similarly, not all forms of subliminal advertising, hypnosis, brainwashing, and so on are inconsistent with moral responsibility for an action. It is only when these activities yield physical mechanisms that are not reasons-responsive that they rule out moral responsibility. Thus, the theory that associates moral responsibility with actual-sequence reasons-responsiveness can help to explain our intuitive distinctions between causal influences that are consistent with moral responsibility and those that are not.

Consider also the class of legal defenses that might be dubbed "Twinkie-type" defenses. This kind of defense claims that an agent ought not to be punished because he ate too much junk food (and that this impaired his capacities, etc.). In the approach presented here, the question of whether an agent ought to be punished is broken into two parts: (1) Is the agent morally responsible (i.e., rationally accessible to punishment), and (2) if so, to what degree ought the agent to be punished? The theory of moral responsibility I have presented allows us to respond positively to the first question in the typical "Twinkie-type" case.

Even if an individual has eaten a diet composed only of junk food, it is highly implausible to think that this yields a biological process that is not weakly reasons-responsive. At the very most, such a process might not be strongly reasons-responsive, but strong reasons-responsiveness is *not* necessary for moral responsibility. Our outrage at the suggestion that a junk food eater is not morally responsible may come from two sources. The outrage could be a reaction to the "philosophical" mistake of demanding strong rather than weak reasons-responsiveness; or the outrage could be a reaction to the implausible suggestion that junk food consumption yields a mechanism that is not weakly reasons-responsive.

Thus the theory of responsibility supports the intuitive idea that Twinkie-type defendants are morally responsible for what they do. Of course, the question of the appropriate *degree* of punishment is a separate question; but it is important to notice that it is *not* a consequence of the theory of responsibility that an agent who acts on a mechanism that is weakly but not strongly reasons-responsive is properly punished to a *lesser* degree than an agent who acts on a mechanism that is strongly reasons-responsive. This may, but need not be, a part of one's full theory of punishment.

Temporal Considerations

I wish to consider a problem for the theory of responsibility that I have been developing. This problem will force a refinement in the theory. Suppose Max (who enjoys drinking but is not an alcoholic) goes to a party where he drinks so much

that he is almost oblivious to his surroundings. In this state of intoxication he gets into his car and tries to drive home. Unfortunately, he runs over a child who is walking in a crosswalk. Although the actual-sequence mechanism issuing in Max's running over the child is plausibly taken to lack reasons-responsiveness, we may nevertheless feel that Max *is* morally responsible for running over the child.

This is one case in a class of cases in which an agent acts at a time T_1 on a reasons-responsive mechanism that causes him to act at T_2 on a mechanism that is *not* reasons-responsive. Further, Max ought to have known that getting drunk at the party would lead to driving in a condition in which he would be unresponsive. Thus, Max can be held morally responsible for his action at T_2 by virtue of the operation of a suitable sort of reasons-responsive mechanism at a prior time T_1. When one acts on a reasons-responsive mechanism at time T_1 and one ought to know that so acting will lead to acting on an unresponsive mechanism at some later time T_2, one can be held morally responsible for so acting at T_2. Thus, the theory of moral responsibility should be interpreted as claiming that moral responsibility for an act at T requires the actual operation of a reasons-responsive mechanism at T or some suitable earlier time. (For simplicity's sake, I suppress mention of the temporal indexation below.)

An individual might cultivate dispositions to act virtuously in certain circumstances. It might even be the case that when he acts virtuously, the motivation to do so is so strong that the mechanism is not reasons-responsive. But insofar as reasons-responsive mechanisms issued in the person's cultivation of the virtue, that person can be held morally responsible for his action. It is only when it is true that at no suitable point along the path to the action did a reasons-responsive mechanism operate that an agent will not properly be held responsible for an action.

Semicompatibilism

I have presented a very sketchy theory of responsibility. The basic idea would have to be developed and explained much more carefully in order to have a fully adequate theory of responsibility, but enough of the theory has been given to draw out some of its implications. My claim is that the theory sketched here leads to compatibilism about moral responsibility and such doctrines as God's foreknowledge and causal determinism.

Let us first consider the relationship between causal determinism and moral responsibility. The theory of moral responsibility presented here helps us to reconcile causal determinism with moral responsibility, even if causal determinism is inconsistent with freedom to do otherwise. The case for the incompatibility of causal determinism and freedom to do otherwise is different from (and stronger than) the case for the incompatibility of causal determinism and moral responsibility.

Causal determinism can be defined as follows:

> *Causal determinism* is the thesis that, for any given time, a complete statement of the facts about the world at that time, together with a complete statement of the laws of nature, entails every truth as to what happens after that time.

Now the "basic argument" for the incompatibility of causal determinism and freedom to do otherwise can be presented. If causal determinism obtains, then (roughly speaking) the past together with the natural laws entail that I act as I do now. So if I am free to do otherwise, then I must either have power over the past or power over the laws of nature. But since the past and the laws of nature are "fixed"—for instance, I cannot now so act that the past would have been different from what it actually was—it follows that I am not now free to do otherwise.[21]

This is obviously a brief presentation of the argument; a more careful and detailed look at the "basic argument" is beyond the scope of this presentation.[22] It should be evident, however, that a compatibilist about causal determinism and freedom to do otherwise must either deny the fixity of the past or the fixity of the laws. That is, such a compatibilist must say that an agent can have it in his power at a time so to act that the past would have been different from what it actually was, or that an agent can have it in his power so to act that a natural law that actually obtains would not obtain.[23] Even if these compatibilist claims are not obviously false, they are certainly not easy to swallow.

The approach to moral responsibility developed here allows us to separate compatibilism about causal determinism and moral responsibility from compatibilism about causal determinism and freedom to do otherwise. The theory says that an agent can be held morally responsible for performing an action insofar as the mechanism actually issuing in the action is reasons-responsive; the agent need not be free to do otherwise. As I explain below, reasons-responsiveness of the actual-sequence mechanism is consistent with causal determination. Thus a compatibilist about determinism and moral responsibility can *accept* the fixity of the past and the fixity of the natural laws. He need not accept the unappealing claims to which the compatibilist about causal determinism and freedom to do otherwise is committed. If it is the "basic argument" that pushes one to incompatibilism about causal determinism and freedom to do otherwise, this need not also push one toward incompatibilism about causal determinism and moral responsibility.

The theory of responsibility requires reasons-responsive mechanisms. For a mechanism to be reasons-responsive, there must be a possible scenario in which the same kind of mechanism operates and the agent does otherwise; but, of course, sameness of kind of mechanism need not require sameness of all details, even down to the "micro" level. Nothing in our intuitive conception of a kind of mechanism leading to action or in our judgments about clear cases of moral responsibility requires us to say that sameness of kind of mechanism implies sameness of micro details. Thus, the scenarios pertinent to the reasons-responsivenes of an actual-sequence mechanism may differ with respect both to the sort of incentives the agent has to do otherwise and the particular details of the mechanism issuing in action. (Note that if causal determinism obtains and I do X, then one sort of mechanism which actually operates is a "causally determined to do X" type of mechanism. But of course this kind of mechanism is not germane to responsibility ascriptions insofar as it is not temporally intrinsic. And whereas the kind, "causally determined," is temporally intrinsic and thus may be germane, it is reasons-responsive.)

If causal determinism is true, then any possible scenario (with the actual natural laws) in which the agent does otherwise at time *T* must differ in *some* respect from the actual scenario prior to *T*. The existence of such possible scenarios is all that is required by the theory of moral responsibility. It is not required that the agent be able to bring about such a scenario (i.e., that the agent have it in his power at *T* so to act that the past, relative to *T*, would have been different from what it actually was). Furthermore, the existence of the required kind of scenarios is compatible with causal determinism.

The actual-sequence reasons-responsiveness theory of moral responsibility thus yields "semicompatibilism": moral responsibility is compatible with causal determinism, even if causal determinism is incompatible with freedom to do otherwise. Compatibilism about determinism and responsibility is compatible with *both* compatibilism and incompatibilism (as well as agnosticism) about determinism and freedom to do otherwise.[24]

Often incompatibilists use the example discussed above of the demonic scientists who directly manipulate one's brain. They then pose a challenge to the compatibilist: In what way is this sort of case *different* from the situation under causal determinism? There is clearly the following similarity: in both the cases of manipulation and determination, conditions entirely "external" to the agent causally suffice to produce an action. Thus, it may be that neither agent is free to do otherwise. However, as I argued above, there seems to be a crucial difference between the case of direct manipulation and "mere" causal determination. In a case of direct manipulation of the brain, it is likely that the process issuing in the action is not reasons-responsive, whereas the fact that a process is causally deterministic does not in itself bear on whether it is reasons-responsive. The force of the incompatibilist's challenge can be seen to come from the plausible idea that in neither case does the agent have freedom to do otherwise; but it can be answered by pointing to a difference in the actual-sequence mechanisms.

The same sort of considerations show that moral responsibility is consistent with God's foreknowledge, even if God's foreknowledge is incompatible with freedom to do otherwise. Let us suppose that God exists and thus knew in the past exactly how I would behave today. If I am free to do otherwise, then I must be free so to act that the past would have been different from what it actually was (i.e., so to act that God would have held a different belief about my behavior from the one he actually held). However, the past is fixed, and so it is plausible to think that I am not free to do otherwise, if God exists.

God's existence, however, is surely compatible with the operation of a reasons-responsive mechanism. God's belief is not a part of the mechanism issuing in my action (on a standard view of the nature of God). His belief is not what causes my action; rather, my action explains his belief. Thus there are possible scenarios in which the actual kind of mechanism operates and issues in my doing otherwise. (In these scenarios, God believes correctly that I will do other than what I do in the actual world.) Again, the cases for the two sorts of incompatibilism—about divine foreknowledge and responsibility and about divine foreknowledge and freedom to do otherwise—are *different*, and the actual-sequence reasons-responsiveness theory yields semicompatibilism.[25]

Structure and History

In this section I wish to contrast my approach to moral responsibility with a class of theories that might be called "mesh" theories of responsibility. My approach is a historical theory.

Consider first a "hierarchical" model of moral responsibility. In this model, a person is morally responsible for an action insofar as there is a mesh between a higher order preference and the first-order preference that actually moves him to action. On one version of this theory, which is suggested by some remarks by Harry Frankfurt, an agent is morally responsible for an action if there is conformity between his "second-order volition" and "will" (the first-order desire that moves the person to action).[26]

In another version of the theory, moral responsibility for an action is associated with conformity between "identification" and will.[27] According to Frankfurt's suggestion, one way of identifying with a first-order desire would be to formulate an unopposed second-order volition to act on it, together with a judgment that no further reflection would cause one to change one's mind.

The problem with such hierarchical "mesh" theories, no matter how they are refined, is that the selected mesh can be produced via responsibility-undermining mechanisms. After all, a demonic neurophysiologist can induce the conformity between the various mental elements via a sort of direct electronic stimulation that is not reasons-responsive. I believe that the problem with the hierarchical mesh theories is precisely that they are purely structural and ahistorical. It matters what kind of process issues in an action. Specifically, the mechanism issuing in the action must be reasons-responsive.

The "multiple-source" mesh theories are also purely structural. Rather than positing a hierarchy of preferences, these theories posit different sources of preferences. One such theory is that of Gary Watson, according to which there are "valuational preferences" (which come from reason) and motivational preferences.[28] Employing Watson's theory, one could say that an agent is morally responsible for an action insofar as there is a mesh between the valuational and motivational preference to perform the action.[29]

Again the problem is that such a theory is purely structural. The mesh between elements of different preference systems may be induced by electronic stimulation, hypnosis, brainwashing, and so on. Moral responsibility is a *historical* phenomenon; it is a matter of the kind of mechanism that issues in action.[30]

Conclusion

I have presented a sketch of a theory that purports to identify the class of actions for which persons are rationally accessible to moral praise and blame, and reward and punishment. I have claimed that this theory captures our clear intuitive judgments about moral responsibility, and that it helps to reconcile moral responsibility with causal determinism. I certainly have not *proved* that moral responsibility is compatible with causal determinism. Rather, my strategy has been to argue that the approach presented here allows the compatibilist about moral responsibility

and determinism to avoid the commitments of the compatibilist about freedom to do otherwise and determinism. There might be other sorts of challenges to compatibilism about determinism and moral responsibility that my approach does not, in itself, answer.

The theory I have presented builds upon and extends the approaches of others. It avoids some of the most pressing objections to similar types of theories. These objections might seem convincing if one fails to "hold fixed" the actual-sequence mechanism, or if one employs strong rather than weak reasons-responsiveness, or if one does not suitably temporally index the theory.

I wish to end with a few suggestions about the relationship between the theory of moral responsibility presented here and punishment. A theory of moral responsibility needs to explain why certain creatures (and not others) are appropriate candidates for punishment. Punishment, of course, involves treating an individual "harshly" in some manner. It affects the desirability of performing a certain action. That is, punishment involves reacting to persons in ways to which the mechanisms on which they act are sensitive. My suggestion is that punishment is appropriate only for a creature who acts on a mechanism "keyed to" the kind of incentives punishment provides.

My point here is not that the justification of punishment is "consequentialist"—that it alters behavior. (Of course, this kind of justification does not in itself distinguish punishment from aversive conditioning.) Indeed, it is metaphysically possible that an individual's total pattern of choices and actions throughout life be "unalterable" by virtue of a continuous string of Frankfurt-type situations. (It is even possible that *no* human's behavior is alterable, because it is possible that all human beings are subject to Frankfurt-type counterfactual interventions.) My justification is nonconsequentialist and "direct": punishment is an appropriate reaction to the actual operation of reasons-responsive mechanisms. When it is justified, punishment involves a kind of "match" between the mechanism that produces behavior and the response to that behavior.

The theory of moral responsibility, then, provides some insight into the appropriateness of punishment for certain actions. But it does not in itself provide a full account of the appropriate *degrees* of punishment. For instance, it may be the case that the appropriate degree of severity of punishment for a particular action is less than (or greater than) the magnitude of the incentive to which the actual-sequence mechanism is responsive. This is entirely compatible with saying that punishment—being a "provider of reasons"—is appropriately directed to agents who act on reasons-responsive mechanisms.

Notes

1. Strawson calls the attitudes involved in moral responsibility the "reactive attitudes": P. F. Strawson, "Freedom and Resentment," *Proceedings of the British Academy* 48 (1962): 1–25.

2. Some contemporary versions of similar theories are found in Alasdair MacIntyre, "Determinism," *Mind* 56 (1957): 28–41; Jonathan Glover, *Responsibility* (New York: Humanities Press, 1970); Herbert Fingarette, *The Meaning of Criminal Insanity* (Berkeley: University of California Press, 1972); Wright Neely, "Freedom and Desire," *Philosophical Review* 83

(1974): 32–54; Timothy Duggan and Bernard Gert, "Free Will as the Ability to Will," *Nous* 13 (1979): 197–217; Lawrence Davis, *A Theory of Action* (Englewood Cliffs, N.J.: Prentice-Hall, 1979); Michael Levin, *Metaphysics and the Mind-Body Problem* (Oxford: Clarendon, 1979); Robert Nozick, *Philosophical Explanations* (Cambridge: Harvard University Press, 1981); and Daniel Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge: MIT Press, 1984). For an excellent survey of some aspects of these approaches, see David Shatz, "Free Will and the Structure of Motivation," in *Midwest Studies in Philosophy* 10, ed. Peter French, Howard Weittstein, and Theodore Uehling (Minneapolis: University of Minnesota Press, 1985, pp. 444–74.

3. I contrast this kind of bank teller with one who, in exactly the same circumstances, does not have an irresistible impulse to comply with the demand. Such a teller may be morally responsible (though not necessarily *blameworthy*) for handing over the money.

4. John Locke presented an interesting example of a man who voluntarily stays in a room which, unknown to him, is locked: John Locke, *Essay Concerning Human Understanding*, Bk. II, chap. 12 Secs. 8–11. For a number of examples of agents who are morally responsible for actions although they could not have done otherwise, see Harry Frankfurt, "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 46, no.23 (1969): 829–39. Also see John Martin Fischer, "Responsibility and Control," *Journal of Philosophy* 79, no. 1 (1982): 24–40.

5. For a vigorous and interesting criticism of this description of the case, see Peter van Inwagen, "Ability and Responsibility," *The Philosophical Review* 87 (1978): 201–24, reprinted in Peter van Inwagen, *An Essay on Free Will* (Oxford: Clarendon, 1983), pp. 161–82. Although it is inappropriate to pursue the details of the debate here, I defend the claim that there are cases in which an agent is morally responsible for performing an action although he couldn't have done otherwise; see Fischer, "Responsibility and Control."

6. I owe this way of describing the Frankfurt-type cases to Sydney Shoemaker.

7. Here I am indebted to Duggan and Gert, "Free Will as the Ability to Will."

8. Ibid.

9. Robert Nozick requires this sort of close contouring of action to value for his notion of "tracking value": see Nozick, *Philosophical Explanations,* pp. 317–62. In this respect, then, Nozick's notion of tracking value corresponds to strong reasons-responsiveness. Nozick claims that an agent who tracks value displays a kind of moral virtue, but he does not claim that tracking value is a necessary condition for moral responsibility.

10. Here I adopt the constraint that the possible worlds pertinent to the weak reasons-responsiveness of the actual-sequence mechanism must have the same *natural laws* as the actual world.

11. Ferdinand Schoeman has brought to my attention a kind of example that threatens my claim that weak reasons-responsiveness is sufficient for moral responsibility. Imagine someone who is apparently insane. This person commits a barbarous act, such as killing a number of persons on the Staten Island Ferry with a saber. And suppose that this individual would have killed the persons under all possible circumstances except one: he would have refrained if he believed that it was Friday and thus a religious holiday. Intuitively, the individual is highly irrational and should not be considered morally responsible, and yet he seems to satisfy the condition of acting from a reasons-responsive mechanism. Weak reasons-responsiveness obtains by virtue of the agent's responsiveness to a bizarre reason, even though the agent is not responsive to a wide array of relevant reasons.

I am aware that this sort of example poses a problem for the theory of responsibility I present here. At this point, I see two possible responses. First, one might claim that in this

kind of case there would be a different mechanism operating in the alternate scenario (in which the agent is responsive) than in the actual sequence. Alternatively, one might restrict the reasons that are pertinent to weak reasons-responsiveness. I hope to discuss such examples and to develop an adequate response in future work.

12. Roughly, one might distinguish between "internalist" and "externalist" accounts of knowledge. An internalist proceeds by requiring that the agent have a certain sort of *justification* for his belief. The externalist abandons the search for refined kinds of justification and requires certain kinds of causal connections between the fact known and the agent's belief.

13. I am obviously presenting only a sketch of a theory of knowledge here. Further, I do not here suppose that this is obviously the *correct* account of knowledge. I am merely pointing to an analogy between my approach to moral responsibility and the externalist conception of knowledge. The approach to knowledge presented here follows those of, among others, Dretske and Nozick: F. Dretske, "Conclusive Reasons," *Australasian Journal of Philosophy* 49 (1971): 1–22; and Nozick, *Philosophical Explanations*, pp. 167–98. Nozick also discusses the analogy between moral responsibility and knowledge.

14. Nozick claims that this fact helps to refute a certain kind of epistemological skeptic. See Nozick, *Philosophical Explanations*, pp. 197–247.

15. Nozick, *Philosophical Explanations*, p.179.

16. Ibid., pp. 180–81. Nozick attributes this example to Avishai Margalit.

17. Ibid., p. 191.

18. I have left extremely vague the crucial notion of "same mechanism." There are certainly very disturbing problems with this notion in epistemology. For a discussion of some of these problems, see Robert Shope, "Cognitive Abilities, Conditionals, and Knowledge: A Response to Nozick," *Journal of Philosophy* 81, no.1 (1984): 29–48. And there may well be similar problems in action theory. Here I am simply relying on some intuitive way of individuating kinds of mechanisms issuing in action, for the purposes of moral responsibility ascriptions. A defense of the sketch of a theory that I am presenting would involve saying more about the individuation of mechanisms.

19. The claim, as stated, relies on the intuition that the physical process *P* is the relevant mechanism. Alternatively, one could simply point out that in Jim's case *there exists* an actually operative mechanism (of kind *P*) that is temporally intrinsic and not reasons-responsive.

20. Daniel Dennett says: "The possibility of short-circuiting or otherwise tampering with an intentional system gives rise to an interesting group of perplexities about the extent of responsibility in cases where there has been manipulation. We are generally absolved of responsibility where we have been manipulated by others, but there is no one principle of innocence by reason of manipulation." Daniel Dennett, "Mechanism of Responsibility," reprinted in *Brainstorms*, ed. Daniel Dennet (Montgomery, Vt.: Bradford, 1978), pp. 233–55, esp. p. 248. My suggestion provides a way of distinguishing responsibility-undermining manipulation from manipulation that is consistent with responsibility.

21. For some contemporary developments of the "basic argument" for incompatibilism, see Carl Ginet, "Might We Have No Choice?" in *Freedom and Determinism*, ed. K. Lehrer (New York: Random House, 1966); David Wiggins, "Towards a Reasonable Libertarianism," in *Essays on Freedom of Action*, ed. T. Honderich (Boston: Routledge and Kegan Paul, 1973); J. W. Lamb, "On a Proof of Incompatibilism," *Philosophical Review* 86 (1977); and Peter van Inwagen, "The Incompatibility of Free Will and Determinism," *Philosophical Studies* 27 (1975), and *An Essay on Free Will* (Oxford: Clarendon, 1983), esp. pp. 55–105.

22. I have discussed the argument in John Martin Fischer, "Incompatibilism," *Philosophical Studies* 43 (1983): 127–37; "Van Inwagen on Free Will," *Philosophical Quarterly* 36 (1986): 252–60; and "Freedom and Miracles," *Nous* 22 (1988), pp. 235–252. For a classic

discussion of the argument, see David Lewis, "Are We Free to Break the Laws?" *Theoria* 47 (1981): 113–21.

23. For an interesting alternative challenge to certain formulations of the "basic argument," see Michael Slote, "Selective Necessity and the Free-Will Problem," *Journal of Philosophy* 82 (1982): 5–24.

24. I believe that Frankfurt is a compatibilistic semicompatibilist. I am an agnostic semicompatibilist, although I am perhaps a latently incompatibilistic semicompatibilist. In "Responsibility and Control" I pointed out that Frankfurt-type cases do not in themselves establish the consistency of causal determinism and moral responsibility. Thus, Frankfurt-type cases leave open the position of "ultra-incompatibilism": Causal determinism is incompatible with moral responsibility, even if moral responsibility does not require freedom to do otherwise. Here I have preferred agnostic (or perhaps incompatibilistic) semicompatibilism to agnostic (or incompatibilistic) ultra-incompatibilism.

25. I have here sketched an approach that attempts to reconcile moral responsibility *for action* with causal determinism and God's foreknowledge. My approach relies on the claim that moral responsibility for an action does not require freedom to do otherwise. Elsewhere I have argued that, whereas an agent can be morally responsible for performing an action although he could not have done otherwise, an agent cannot be held responsible for *not* performing an action he could not have performed: John Martin Fischer, "Responsibility and Failure," *Proceedings of the Aristotelian Society* 86 (1985–86): 251–70. If this "asymmetry thesis" is true, then I still have not reconciled moral responsibility *for omissions* (or perhaps, for "not-doings") with causal determinism (and divine foreknowledge).

I do not have the space here fully to develop my theory of responsibility for not performing actions. But I can say that, even if an agent is not responsible for failing to do something he could not do, an agent may be held morally responsible for *something* (perhaps, a "positive" action). And so he will be accessible to praise or blame. I believe that such a theory of moral responsibility can be developed so as to reconcile causal determinism (and divine foreknowledge) with the moral attitudes we think are intuitively appropriate.

26. Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68 (1971): 5–20, esp. p. 15.

27. Frankfurt discusses the notion of identification in "Identification and Externality," in *The Identities of Persons*, ed. A. O. Rorty. (Berkeley: University of California Press, 1976); and "Identification and Wholeheartedness," chap. 2 of *Responsibility, Character, and the Emotions*, ed. Ferdinand Schoeman (Cambridge: Cambridge University Press, 1987).

28. Gary Watson, "Free Agency," *Journal of Philosophy* 72 (1975): 205–20.

29. I am not sure whether Watson himself is committed to the sufficiency of the mesh for moral responsibility. He is committed to the claim that an agent is free insofar as he has the power to effect a mesh between the valuational and motivational systems. Ibid., p. 216.

30. Moral responsibility is in this respect like such notions as justice and love for a particular person. Nozick argues in *Anarchy, State, and Utopia* (New York: Basic Books, 1974) that justice and love are historical rather than "current time-slice" notions. Purely structural approaches to moral responsibility are inadequate in a way that is parallel to the inadequacy of current time-slice approaches to justice.