Cambridge Studies in Philosophy and Law

GENERAL EDITOR: Gerald Postema
(University of North Carolina, Chapel Hill)

ADVISORY BOARD

Jules Coleman (Yale Law School)
Anthony Duff (University of Stirling)
David Lyons (Boston University)
Neil MacCormick (University of Edinburgh)
Stephen Munzer (U.C.L.A. Law School)
Phillip Pettit (Australian National University)
Joseph Raz (University of Oxford)
Jeremy Waldron (Columbia University)

Other books in the series

Jeffrie G. Murphy and Jean Hampton: *Forgiveness and mercy*
Stephen R. Munzer: *A theory of property*
R. G. Frey and Christopher W. Morris (eds.): *Liability and responsibility: Essays in law and morals*
Robert F. Schopp: *Automatism, insanity, and the psychology of criminal responsibility*
Steven J. Burton: *Judging in good faith*
Jules Coleman: *Risks and wrongs*
Suzanne Uniacke: *Permissible killing: The self-defense justification of homicide*
Jules Coleman and Allen Buchanan (eds.): *In harm's way: Essays in honor of Joel Feinberg*
Warren F. Schwartz (ed.): *Justice in immigration*
R. A. Duff (ed.): *Philosophy and the criminal law*
Larry Alexander (ed.): *Constitutionalism*
R. Schopp: *Justification defenses and just convictions*
Anthony Sebok: *Legal positivism in American jurisprudence*
William Edmundson: *Three anarchical fallacies: An essay on political authority*
Arthur Ripstein: *Equality, responsibility, and the law*
Heidi M. Hurd: *Moral combat*

Responsibility and Control

A Theory of Moral Responsibility

JOHN MARTIN FISCHER
UNIVERSITY OF CALIFORNIA, RIVERSIDE

MARK RAVIZZA, S.J.
JESUIT SCHOOL OF THEOLOGY, BERKELEY

CAMBRIDGE
UNIVERSITY PRESS

regulative control. If this is indeed so, then a line of argument opens that has some chance of answering at least some of the skeptic's challenges to our moral responsibility.

This line can be sketched as follows. Guidance control, and not regulative control, is the control that is associated with moral responsibility; that is, guidance control in itself (and apart from regulative control) satisfies the freedom-relevant condition of moral responsibility. If this is correct, then the indirect challenge to our moral responsibility (based on the possible truth of causal determinism) can be *sidestepped*. The indirect challenge proceeds by contending that causal determinism rules out alternative possibilities; but if alternative possibilities are not required for moral responsibility, the indirect challenge is rendered irrelevant.

III. GUIDANCE CONTROL

It certainly appears that the Frankfurt-type examples sketched here are consistent with causal determinism. That is, it is possible that in the two examples, Sally's guidance of her car and Sam's shooting the mayor, causal determinism obtains. And if the relevant sort of control is indeed present in the examples – guidance control – this suggests that guidance control is compatible with causal determinism. We contend that this is so – that guidance control, the sort of control that grounds moral responsibility, is compatible with causal determinism. We wish now to defend this contention by beginning to provide a more explicit account of guidance control, according to which guidance control can be seen to be compatible with causal determinism. This account will provide part of a strong prima facie case for the compatibility of guidance control and thus moral responsibility with causal determinism.[6]

Recall that we shall generally proceed by seeking to establish a wide reflective equilibrium within the domain of phenomena associated with moral responsibility. Thus, an account of moral responsibility should capture our intuitive judgments about clear cases. In order to generate a principle that might underlie our reactions to relatively clear cases, it is useful to begin by considering examples in

[6] Later in the book (in Chapter 6) we will consider an argument that causal determinism rules out moral responsibility *directly* (i.e., not in virtue of ruling out alternative possibilities); it will not be until we refute this argument that the case for the compatibility of moral responsibility and causal determinism will have been made more decisively.

which we are inclined to think that an agent cannot legitimately be held morally responsible for what he does.

Imagine that Jones has been hypnotized. The hypnotist has induced an urge that will impel Jones to punch the nearest person after hearing the telephone ring. Insofar as Jones did not consent to this sort of hypnotic suggestion (perhaps he has undergone hypnosis to help him stop smoking), it seems unreasonable to say that he has guidance control of his 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, which causes Smith to murder his neighbor. Because of the nature of the causal history of the action, it is apparent that Smith does not control his behavior in the relevant sense.

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 insofar as they lack the relevant sort of control. Imagine, for instance, that neurophysiologists of the future can isolate certain key parts of the brain that 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 believe that the agent has guidance control of his behavior. Thus, we do not hold him morally responsible for it. Certain sorts of mental disorders – extreme phobias, for instance – may also issue in behavior that the agent does not control in the relevant sense.

Many people believe that there can be genuinely "irresistible" psychological impulses. If so, then these may issue in behavior the agent does not control. 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 that they are not morally responsible for acquiring the addiction in the first place).

Also, certain coercive threats (and perhaps offers) rule out moral responsibility. The bank teller who is told that he will be shot unless he hands over the money might have an overwhelming and irresistible desire to comply with the threat. Insofar as he acts from such an impulse, it is plausible to suppose that he does not have guidance control of his action.[7]

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. These factors issuing in behavior are, intuitively, responsibility-undermining factors.

We can contrast such cases – in which some responsibility-undermining factor actually 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 forth, we think that 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. In such a case, we feel confident in ascribing guidance control to you.

On first consideration of this array of cases, it might be thought that there is a fairly obvious way of distinguishing the clear cases of moral responsibility from the clear cases of a 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 a 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 (and thus regulative control).

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

[7] Contrast this kind of bank teller with one who, in exactly the same circumstances, does not have an irresistible impulse to comply with the threat, but simply complies with the threat because this is the reasonable and prudent thing to do. Such a teller might be morally responsible (though not necessarily blameworthy) for handing over the money.

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 – his 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 agents who are morally responsible and those who are not consists in the "reasons-responsiveness" of the agents (and thus their possession of regulative control).

But of course we have already seen that there are cases in which an agent can be held morally responsible for performing an action, even though he couldn't have done otherwise and is *not* "reasons-responsive": the Frankfurt-type cases. 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. That is, in a Frankfurt-type case, no responsibility-undermining factor operates in the actual sequence, but such a factor operates in the alternative sequence. As we have argued, in a Frankfurt-type case the agent has guidance control of his action, even though he lacks regulative control.

We believe, then, that the Frankfurt-type cases invite us to look more carefully at the characteristics of the *actual sequence that leads to the action*. That is, these cases invite us to develop what we shall call an "actual-sequence" account of moral responsibility. By an "actual-sequence" approach, we mean an approach to moral responsibility that does *not* require alternative possibilities. In contrast to traditional views, an actual-sequence model of moral responsibility holds that ascriptions of responsibility do *not* depend on whether agents are free to pursue alternative courses of action (and thus have alternative scenarios genuinely accessible to them); rather, what is important is (roughly speaking) what the agents actually do, and how their actions come to be performed.

Frankfurt-type examples underscore the importance of distinguishing what happens in the *actual* sequence of events from what

happens in some *alternative* scenario (or range of alternative scenarios). In these examples, the factor that would undermine an agent's responsibility – for example, the direct manipulation of the brain – only occurs in the alternative sequence(s). In the actual course of events, no responsibility-undermining factor operates: the agent chooses freely, and acts in accordance with his choice, in just the way he would have, had there been no "counterfactual intervener" at all. The Frankfurt-type examples highlight the fact that, as long as no responsibility-undermining factor actually operates, an agent may be morally responsible, even though such a factor would have played a role in the alternative scenario (and thus the agent lacks alternative possibilities).

We contend that one very useful way to develop an actual-sequence approach to moral responsibility is to switch from a focus on the relevant *agents* and their properties, to a focus on the processes or "mechanisms" that actually lead to the action. In other words, we infer from the Frankfurt-type cases (among other things) that it is better to take what might be called a "mechanism-based" approach to moral responsibility than an "agent-based" approach.

As we pointed out, in a Frankfurt-type case the agent could not have done otherwise, and thus the *agent* is not reasons-responsive. But it is crucial to see that in these cases the kind of mechanism that *actually* operates *is* reasons-responsive, even though the kind of mechanism that *would* operate – that is, that does operate in the alternative scenario – is *not* reasons-responsive. Note that, although we employ the term "mechanism," we do *not* mean to point to anything over and above the process that leads to the relevant upshot; instead of talking about the mechanism that leads to (say) an action, we could instead talk about the process that leads to the action, or the "way the action comes about."[8] In the Frankfurt-type case, "Assassin" (in which Sam shoots the mayor on his own and Jack does not actually intercede), Sam'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 stimula-

[8] Thus, we are *not* committed to any sort of "reification" of the mechanism; that is, we are not envisaging a mechanism as like a mechanical object of some sort. The mechanism leading to an action is, intuitively, the way the action comes about; and, clearly, actions can come about in importantly *different ways*.

tion of Sam's brain.[9] (Recall that, had Sam been inclined to waver and not shoot the mayor, Jack's device would have been triggered, and it would have stimulated Sam's brain so as to ensure that he would choose to shoot the mayor and in fact shoot the mayor.) And this alternative-sequence mechanism is not reasons-responsive. Thus, the actual-sequence *mechanism* can be reasons-responsive, even though the *agent* is not reasons-responsive. (*He* couldn't have done otherwise.) The switch from a focus on the responsiveness profiles of *agents* to those of *mechanisms* is important; we explain and develop this move further in the appendix to this chapter.[10]

We believe that there is a further interesting – and important – feature of Frankfurt-type examples. In a Frankfurt-type case the actually operative mechanism is in some important sense the "agent's own," whereas the mechanism that *would have operated* (i.e., that operates in the alternative scenario) is not. For example, if Jack had intervened and electronically stimulated Sam's brain so as to cause him to shoot the mayor, this mechanism would *not* have been Sam's own (in some intuitive sense). As we shall explain later in the book (especially in Chapter 8), the two dimensions of assessment – whether a mechanism is the agent's own and its degree of reasons-responsiveness – appear to be conceptually *distinct*. We shall treat them as *two separate dimensions* of guidance control.

One might then employ the following condition as part of a theory that distinguishes the relatively clear cases of moral responsibility from cases of a lack of it: an agent exhibits guidance control of an action insofar as the mechanism that actually issues in the action is his own, reasons-responsive mechanism. (Later we shall revise this condition, but it is a useful starting point.) In this (and the following) chapter, we shall focus on the second element: reasons-responsiveness. Later in the book we shall return to the important idea

[9] Alternatively, one could say that in the actual and alternative sequences the action comes about in importantly different ways.

[10] Arthur Koestler employs "mechanism of thought" to refer to a process of thought, and he talks about how this sort of mechanism can be altered significantly: "Rubashov wondered what other surprises his mental apparatus held in store for him. He knew from experience that confrontation with death always altered the mechanism of thought and caused the most surprising reactions – like the movements of a compass brought close to the magnetic pole" (Arthur Koestler, *Darkness at Noon*, trans. Daphne Hardy [New York: Macmillan, 1941], p. 50). We are indebted to Jerry Burke for bringing this passage to our attention.

captured in the first element: that the mechanism must be the *agent's own*.

Clearly, on the approach to moral responsibility we have begun to sketch, it is crucial to distinguish between the kind of mechanism that operates in the actual sequence and the kind of mechanism that operates in the alternative sequence (or sequences). We must confess that we do not have any general way of specifying when two kinds of mechanisms are the same. This is a potential problem for our approach; it will have to be considered carefully by the reader. But rather than attempting to say much by way of giving an account of mechanism-individuation, we shall simply rely on the fact that people have intuitions about fairly clear cases of "same kind of mechanism" and "different kind of mechanism." For example, we rely on the intuitive judgment that the normal mechanism of practical reasoning is different from deliberations that are induced by significant direct electronic manipulation of the brain, hypnosis, subliminal advertising, and so forth. We believe that the development and application of our approach to moral responsibility will rely primarily on relatively clear intuitions about sameness (and difference) of mechanisms. Given these relatively clear intuitive judgments, the approach should be judged by its fruitfulness in sorting through and illuminating the puzzling and difficult problems to which it is applied.[11]

So far, we have pointed to some cases in which it is intuitively clear that a person lacks guidance control of his actions and thus cannot be held morally responsible for what he has done, and we have also indicated other cases in which it is intuitively clear that an agent has such control and thus can be held responsible for his actions. We have suggested a principle that might help to distinguish the two kinds of cases. In addition to the important notion of a mechanism's being the agent's own, the principle employs two salient ingredients: reasons-responsiveness and the distinction between actual-sequence and alternative-sequence mechanisms. We

[11] More specifically, the idea is that we shall attempt to employ relatively clear judgments about mechanism individuation as part of a general theory of moral responsibility which will systematize our reflective, considered judgments – and render them coherent even in problematic contexts. Note, also, that if the cases in which people are unclear about mechanism individuation are also cases in which they are unclear about moral responsibility, then our theory will "capture" or "mirror" the pretheoretic fuzziness. Although our goal is to help to guide reflective individuals in their judgments about control and moral responsibility, it may be that there is some genuine indeterminacy in the phenomena our theory purports to systematize.

now wish to explain these ingredients more carefully, beginning with the notion of reasons-responsiveness; we shall return to the distinction between actual-sequence and alternative-sequence mechanisms in the appendix to this chapter. (In our discussions of these ideas, we shall sometimes suppress mention of the condition that the actually operative mechanism must be the *agent's own;* this is solely for simplicity's sake.)

IV. REASONS-RESPONSIVENESS

We shall discuss (in this chapter) two kinds of reasons-responsiveness: strong and weak. Begin with strong reasons-responsiveness. Suppose that a certain kind *K* of mechanism actually issues in an action. Strong reasons-responsiveness obtains under the following conditions: if *K* were to operate and there were sufficient reason to do otherwise, the agent would *recognize* the sufficient reason to do otherwise and thus *choose* to do otherwise and *do* otherwise. In other words, under circumstances in which the actual kind of mechanism operates and there are sufficient reasons for the agent to do otherwise, three conditions must be satisfied (in order for there to be strong reasons-responsiveness): the agent must *take* the reasons to be sufficient, *choose* in accordance with the sufficient reasons, 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 reasons and his choice, and in the connection between choice and action.

The first kind of failure is a failure to be receptive to reasons. Here there are sufficient reasons (say) to perform some action, but the agent does not recognize these reasons. When this sort of failure is due to an inability, it is typically associated with delusional psychosis.[12] The second kind of failure is a failure of reactivity – a failure to be appropriately affected by beliefs. Here the agent recognizes certain reasons as sufficient (say) for performing an action, but he does not choose in accordance with this recognition.[13] Lack of reac-

[12] Here we are indebted to Timothy Duggan and Bernard Gert, "Free Will As the Ability to Will," *Noûs* 13 (1979), pp. 197–217; reprinted in Fischer, ed., *Moral Responsibility.*

[13] This throws into clear relief the fact that we are using "sufficient reason" in the sense of "justificatorily sufficient reason," rather than that of "motivationally sufficient reason." The point is that an agent may acknowledge that a reason

tivity afflicts certain compulsive or phobic neurotics.[14] Also, this sort of failure may result from weakness of will. Finally, there is the failure successfully to translate one's choice into action; this sort of failure may reflect various kinds of physical incapacities, or even (again) weakness of the will. 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 his choice, and his choice and action. The agent's actions would fit the contours of reasons closely: he would be tightly aligned with reasons. Robert Nozick requires this sort of close contouring of action to value for his notion of "tracking value."[15] 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.

Whereas such a close contouring of actions to reasons is no doubt desirable in many respects, we do not believe that strong reasons-responsiveness is a necessary condition for guidance control and moral responsibility. To see this, imagine that as a result of the unimpaired operation of the normal human faculty of practical reasoning Jennifer decides to go (and goes) to the basketball game tonight, and that she has sufficient reason to do so. But suppose that she 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 she had a strict deadline for an important manuscript (which she couldn't meet, if she were to go to the game). She nevertheless would have chosen to go to the game, even though she would have recognized that she had sufficient reason to stay home and work. It seems that Jennifer actually goes to the basketball game freely and can reasonably be held morally responsible for going; and yet the actual-sequence mechanism which results in her action is not reasons-responsive in the strong sense. The failure of strong reasons-responsiveness here stems from Jennifer's disposition toward weakness of the will.

justifies a certain course of action – it is, all things considered, his strongest or best reason for action – without being *motivated* by that reason.

14 See Duggan and Gert, "Free Will As the Ability to Will."

15 Robert Nozick, *Philosophical Explanations* (Cambridge, Mass.: Harvard University Press, 1981), pp. 317–362.

Going to the basketball game is plausibly thought to be a *morally neutral* act; on 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 possibility of weakness of will also shows that strong reasons-responsiveness is not necessary for moral responsibility for *commendable* acts. Suppose, for example, that Jennifer devotes her afternoon to working for the United Way (and her decision and action proceed via what most people would take to be a "responsibility-conferring mechanism"). And imagine that, if she had a sufficient reason to refrain, it would (again) have been her publication deadline. But imagine that she would have devoted her time to charity, even if she had had such a reason not to. Here it seems that Jennifer is both morally responsible and praiseworthy for doing what she does, 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 Leonard steals a book from a store, knowing full well that it is morally wrong for him to do so and that he will be apprehended and thus that it is not prudent of him to do so. Nevertheless, the actual sequence may be responsibility-conferring; no factors that intuitively undermine moral responsibility may actually operate. (Of course, we assume here 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 Leonard is morally responsible for stealing the book, but his actual-sequence mechanism is not strongly reasons-responsive: there actually is sufficient reason (both moral and prudential) to do otherwise, and yet he steals the book.

All three cases presented here provide problems for the claim that strong reasons-responsiveness is necessary for moral responsibility. Strong reasons-responsiveness may be both necessary and sufficient 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. But, of course, not all agents who are morally responsible are morally commendable (or even maximally prudent). We believe that it is useful to explore the idea that moral responsibility requires only a looser kind of fit between reasons and action: "weak reasons-responsiveness." We shall adopt this idea as a working hypothesis in this chapter; in the following chapter, we shall develop some refinements.

It is perhaps easiest to understand weak reasons-responsiveness by contrasting this view with strong reasons-responsiveness. Under the requirement of strong reasons-responsiveness, we hold fixed the actual kind of mechanism and ask what would happen, if there were a sufficient reason to do otherwise.[16] In contrast, under weak reasons-responsiveness, we (again) hold fixed the actual kind of mechanism, and we then simply require that there exist *some* possible scenario (or possible world) in which there is a sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise.[17] (We shall adopt the constraint that the possible worlds pertinent to the responsiveness of the actual-sequence mechanism must have the same natural laws as the actual world.)

16 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), which are *most similar* to the actual world. (Perhaps there is just one such world, or perhaps there is a sphere of many such worlds.)

This follows from the fact that strong reasons-responsiveness is defined in terms of a *subjunctive conditional*, given the possible-worlds semantics for such conditionals. The relevant conditional (defining strong reasons-responsiveness) is something like this: if the actual kind of mechanism were to operate and the agent were to have a sufficient reason to do otherwise, he would do otherwise. On the possible worlds semantics for this sort of conditional, it is true just in case the consequent is true in the possible-world(s) most similar to the actual world in which the antecedent is true. That is, the conditional is true just in case the agent would do otherwise in the possible worlds most similar to the actual world in which the actual kind of mechanism operates and the agent has a sufficient reason to do otherwise. For developments of the possible-worlds semantics for subjunctive (and counterfactual) conditionals, see Robert Stalnaker, "A Theory of Conditionals," in Nicholas Rescher, ed., *Studies in Logical Theory, American Philosophical Quarterly* Series, vol. 1 (Oxford: Blackwell, 1968), pp. 98–112; and David Lewis, *Counterfactuals* (Cambridge, Mass.: Harvard University Press, 1973).

Note that strong reasons-responsiveness is defined in terms of the subjunctive conditional, *not* entailment. That is, it is not supposed that strong reasons-responsiveness requires that in *any* possible world in which the actual kind of mechanism operates and the agent has a sufficient reason to do otherwise, he does otherwise. This condition corresponds to an *entailment:* the actual mechanism's operating and the agent's having a sufficient reason to do otherwise *entails* his doing otherwise. This is not strong, but "*super* reasons-responsiveness."

17 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.

Consider, again, Jennifer's decision to go to the basketball game. In this situation, if she were to have a sufficient reason to do otherwise, it would be a publication deadline. And she would under such circumstances be weak-willed and still go to the game. But certainly there exists *some* scenario in which the actual mechanism operates, she has sufficient reason not to go to the game, and she doesn't go. Suppose, for instance, that Jennifer is told that she will have to pay one thousand dollars for a ticket to the game. In this situation, she presumably would not go to the game. Even though Jennifer is disposed to be weak-willed under some circumstances, there are other circumstances in which she would respond appropriately to sufficient reasons. These are circumstances in which the reasons are considerably *stronger* than the reasons that would exist if she were to have sufficient reason to do otherwise.

Consider, similarly, Jennifer's commendable act of working this afternoon for the United Way. Even though she would do so anyway if she had a publication deadline, she certainly would *not* work for the United Way if to do so she would have to sacrifice her job. Thus, the actual mechanism issuing in her action is weakly reasons-responsive. Also, when an agent wrongly (and imprudently) steals a book (i.e., there actually are sufficient reasons – moral and prudential – not to), his actual mechanism might be responsive to at least *some* possible incentive not to steal. Even an agent who acts against good reasons can be responsive to *some* reasons.

It is reasonable to think that the agent's actual-sequence mechanism *must* be weakly reasons-responsive if he is to have the sort of control required for moral responsibility. If (given the operation of the actual kind of mechanism) he would persist in stealing the book, even if he knew 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 him morally responsible for his action. Arguably, this is because the agent here would not be exhibiting genuine control of his action.

So weak reasons-responsiveness is necessary for moral responsibility. It also seems plausible that weak reasons-responsiveness is *sufficient* for moral responsibility (given that the epistemic conditions are satisfied). That is, it is reasonable to think that Jennifer is morally responsible for going to the basketball game to the extent that she acts on a weakly reasons-responsive mechanism. Similarly, insofar as Leonard's actual-sequence mechanism is at least weakly reasons-

responsive, it seems that he can be held morally responsible for stealing the book.

We have suggested, then, that weak reasons-responsiveness is *all* the responsiveness that is required for the sort of control involved in moral responsibility (given that the relevant mechanism is "the agent's own"). We believe that this suggestion has considerable plausibility, at least to a first approximation. It is then the working hypothesis of this chapter that, on the assumption that the ownership condition has been satisfied (i.e., that the relevant mechanism is the agent's own), weak reasons-responsiveness is necessary and sufficient for the freedom-relevant condition of moral responsibility – guidance control.

V. A BIT ABOUT MECHANISMS

We have suggested that an agent has guidance control of an action insofar as the mechanism that actually issues in the action is the agent's own, reasons-responsive mechanism. But given that various different mechanisms may actually operate in a given case, which mechanism is the one that is relevant?

Suppose that you deliberate (in the normal way) about whether to donate 5 percent of your paycheck to the United Way, and that you decide to make the donation and act on your decision. We might fill in the story so that it is intuitively a paradigmatic case in which you are morally responsible for your 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 you would give 5 percent of your paycheck to the United Way in any logically possible scenario. Thus, this kind of actually operative mechanism is not reasons-responsive.

But presumably a mechanism such as "deliberation prior to giving 5 percent of one's salary to the United Way" is not 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.[18]

[18] A temporally extrinsic or relational mechanism is in this respect similar to temporally relational properties and facts; see Fischer, *Metaphysics*, pp. 111–130.

Note that the operation of a mechanism of the kind "deliberation 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 ascriptions of guidance control (and moral responsibility) that it be a "temporally intrinsic" or "nonrelational" mechanism in the following sense: if a mechanism M issues in act A, then M is relevant to the agent's guidance control of A only if M's operating does not entail that A occurs. We believe that the requirement that a mechanism be temporally intrinsic is an intuitively natural and unobjectionable one. But, 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.[19] Which mechanism is "the" mechanism pertinent to guidance control?

We cannot specify in a general way how to determine which mechanism is "the" mechanism that is relevant to assessment of responsibility. It is simply a presupposition of this theory as presented here that for each act, there is an intuitively natural mechanism that is appropriately selected as the mechanism that issues in action, for the purposes of assessing guidance control and moral responsibility. The problem here is, of course, similar to that of "generalization" theories in ethics. On such an approach, an act is (say) wrong if there would be (for example) certain bad consequences of actions of *that type* generally being done (or the general acceptance of a rule specifying the permissibility of *that type* of action, and so forth). On these approaches, it is assumed that there is some natural, unproblematic way of selecting the relevant general "type" by reference to which the act is to be assessed. A similar assumption lies behind the theory of moral responsibility.

[19] It would seem to be arbitrary to suppose that there is just one mechanism that actually issues in a given action. It would seem similarly arbitrary to suppose that, in moving one's body in a certain way, one is performing just one action. On various plausible views of the nature of action, one may at any time be performing a large number of actions; it may be, however, that only a small number are important or relevant, given the context. Presumably, also, at any given time an individual may possess a large number of properties; but, again, it may be that only a relatively small number of these are important or relevant, given the context. So, also, with mechanisms.

This is one case in a class of cases in which an agent's act at a time *T1* issues from a reasons-responsive sequence, and this act causes his act at *T2* to issue from a mechanism that is not reasons-responsive. Further, Max can reasonably be expected to have known that his getting drunk at the party would lead to his driving in a condition in which he would be unresponsive. Thus, Max can be held morally responsible for his action at *T2* in virtue of the operation of a suitable sort of reasons-responsive mechanism at a prior time *T1*. When one acts from a reasons-responsive mechanism at time *T1*, and one can reasonably be expected to know that so acting will (or may) lead to acting from an unresponsive mechanism at some later time *T2*, one can be held responsible for so acting at *T2*.[21] This sort of case illustrates *one* way in which a prior action on a responsive mechanism can be the basis of an agent's subsequent moral responsibility. In general, 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.

An individual might cultivate dispositions to act virtuously in certain circumstances. It might even be the case that when he acts virtuously, his motivation to do so is so strong that the mechanism is not reasons-responsive. But insofar as reasons-responsive sequences issued in his cultivation of the virtue, he 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 sequence occur that an agent will not properly be held responsible for his action.

This is, admittedly, a sketchy and incomplete treatment of difficult issues; we hope that enough of the flavor of the account has been given for it to be useful for our purposes here. The general approach we are developing is a "tracing" approach: when an agent is morally responsible for an action that issues from a mechanism that is not appropriately reasons-responsive, we must be able to trace back along the history of the action to a point (*suitably related to the action*)

[21] This account will have to remain vague (in various ways). In some contexts, it seems appropriate to hold an agent responsible for a later action (or omission or consequence) that is extremely unlikely to occur, whereas in other contexts the extreme unlikelihood of (say) the action seems to rule out responsibility. This makes it reasonable to think that a full and explicit tracing approach would not simply specify a degree of likelihood that is always employed straightforwardly to ascertain responsibility; rather, the degree of likelihood employed by the tracing approach would need to be context-relative.

where there was indeed an appropriately reasons-responsive mechanism.

VIII. SEMICOMPATIBILISM

We have presented a very sketchy account of guidance control. We propose further to elaborate and refine it in the following chapter (and also later chapters). But enough of the theory has been given to draw out some of its implications. Our claim is that the account sketched here leads to *compatibilism* about moral responsibility and the doctrine of causal determinism.

Let us then consider the relationship between causal determinism and moral responsibility for actions in light of the theory of moral responsibility for actions that we have sketched. The account of guidance control presented here helps us to reconcile causal determinism with moral responsibility for actions, even if causal determinism is inconsistent with freedom to do otherwise. We shall contend that 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 for actions.

The approach to moral responsibility developed here says that an agent can be held morally responsible for performing an action insofar as the mechanism actually issuing in the action is the agent's own, weakly reasons-responsive mechanism; the agent need not be free to do otherwise. And (as we shall explain) reasons-responsiveness of the actual sequence leading to action is consistent with causal determination. Thus a compatibilist about determinism and moral responsibility need not reject any of the very plausible ingredients of the indirect challenges from causal determinism to moral responsibility (presented in Chapter 1). That is, such a compatibilist need not reject such plausible principles as the Principle of the Fixity of the Past, the Principle of the Fixity of the Laws, and the Transfer Principle. If it is the thrust of this set of challenges 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 for actions.

The account of guidance control (and thus responsibility) requires weakly reasons-responsive mechanisms. For a mechanism to be weakly 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 "microlevel," just as nothing in "same kind of house" or "same kind of smile" requires sameness of all details. 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 microdetails. Thus, the scenarios pertinent to the reasons-responsiveness of an actual-sequence mechanism may differ with respect both to the sort of incentives the agent has to do otherwise and to the particular details of the mechanism issuing in action. Note that if causal determinism obtains and I do *A*, then one sort of mechanism that actually operates is a "causally determined to do *A*" 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. Further, there is no plausibility to the suggestion that *all conditions in the past* – no matter how remote or irrelevant – must be included as part of the "mechanism that issues in action."[22]

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 our theory of moral responsibility. It is crucial to our approach that it does *not* require that the agent be able to bring about such a scenario, that is, that he 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. And the *existence* of the required kind of scenarios is surely compatible with causal determinism. Thus, our approach to moral responsibility makes room for responsibility for actions even in a causally deterministic world. The actual-sequence reasons-responsiveness account of guidance control (and moral responsibility) thus helps to yield

[22] We are seeking to capture faithfully our considered judgments about clear cases of moral responsibility (and the lack of it). In doing so, we have employed the ingredient, "same kind of mechanism." We claim that the goal of capturing our considered judgments about clear cases requires us *not* to take a stringent view of "same kind of mechanism" (according to which sameness requires sameness down to microdetails). Of course, if one has some *prior commitment* to the view that causal determinism is incompatible with moral responsibility, then one will be inclined to press for such an interpretation. But, apart from such a commitment, we do not see why one would be inclined toward this view of "same kind of mechanism."

"semicompatibilism": moral responsibility is compatible with causal determinism, even if causal determinism is incompatible with freedom to do otherwise. That is to say, the first step toward semicompatibilism has been taken: the step that pertains to moral responsibility for *actions*.[23]

On our approach, moral responsibility does not require alternative possibilities. Rather, we have an "actual-sequence" approach to moral responsibility. By this we mean (in part) that one should focus on the properties of the actual sequence in making ascriptions of moral responsibility. And these properties are *not* relevant in virtue of pointing to the existence of alternative possibilities – they are relevant to ascriptions of moral responsibility *more directly*.

Notice, however, that these "actual-sequence" properties may indeed be *dispositional* or *modal* properties; as such, their proper analysis may involve reference to other possible scenarios or worlds. That is, we have argued that a certain sort of reasons-responsiveness is required for moral responsibility; we then have analyzed this sort of responsiveness in terms of other possible worlds. Thus, we have associated moral responsibility with a dispositional or modal property. It is important to see that, whereas other possible worlds are relevant to ascertaining whether there is some actually operative dispositional feature (such as weak reasons-responsiveness), such worlds are *not* relevant in virtue of bearing on the question of whether some alternative sequence is *genuinely accessible* to the agent.

On our approach to moral responsibility, then, other possible scenarios are relevant to the issue of whether the *actual sequence* has certain features (such as weak reasons-responsiveness). But it does *not* follow that our approach is committed to the claim that agents can have it in their power to *actualize* such scenarios – that is a quite different matter. Since we do *not* hold that moral responsibility requires alternative possibilities, we do *not* need to say that agents can have it in their power to actualize scenarios different from the actual scenario. And thus we need *not* deny (for example) the basic idea of the second version of the Indirect Challenge to moral responsibility (presented in Chapter 1).

To see this, recall that our discussion of the second version of the Indirect Challenge showed how, if causal determinism is true, one could not even in principle trace out a path along which the natural

[23] We assume here that satisfaction of the ownership condition is compatible with causal determinism. We shall argue for this in Chapter 8.

laws obtain from the actual path to some alternative action. But our approach to moral responsibility does *not* require that agents be able to do otherwise (and thus actualize alternative scenarios); thus, we need *not* run afoul of the plausible idea that our freedom must be the freedom to extend the actual past. By adopting an actual-sequence approach to moral responsibility, we can thus avoid the thrust of the Indirect Challenges. Indeed, this is the great "payoff" of adopting an actual-sequence approach to moral responsibility.

IX. CONCLUSION

In Chapter 1 we laid out two conflicting "tendencies." On the one hand, there is the strong and natural commitment to personhood and moral responsibility, which appears unshakable in the light of the truth (or even the falsity, in certain ways) of causal determinism. On the other hand, there are powerful challenges to this strong and natural belief based precisely on the possible truth of causal determinism.

In this chapter we have taken the first step toward protecting our moral responsibility against at least some of the challenges (especially the Indirect Challenge from causal determinism). First, we argued (based on the Frankfurt-type examples) that the sort of control that traditionally has been deemed necessary for moral responsibility – regulative control – is in fact *not* necessary for moral responsibility. Rather, it is *guidance control* that grounds moral responsibility for actions. Granted, regulative control may typically accompany guidance control; but it is guidance control (and not regulative control) that is the basis of moral responsibility for actions.

Further, reflection on the Frankfurt-type examples leads to the beginnings of an account of guidance control. An agent exercises guidance control of an action to the extent that the action issues from the agent's own, reasons-responsive mechanism. The important notion of a mechanism's being an "agent's own" will be the focus of attention later in the book (particularly in Chapter 8). In this chapter we have begun to understand reasons-responsiveness in terms of the weak reasons-responsiveness of the mechanism leading to the action, and we offered a refinement that allows for tracing back into the past in search of the mechanism relevant to moral responsibility.

Since the Indirect Challenge from causal determinism to our moral responsibility proceeds by arguing that causal determinism rules out regulative control, the realization that regulative control is

not required for moral responsibility goes some distance toward assuaging the worries about causal determinism. In this chapter we also have begun to sketch an account of guidance control of actions according to which such control is compatible with causal determinism. If guidance control is the control associated with moral responsibility, and if it can be shown that it is plausible that guidance control is compatible with causal determinism, then a strong prima facie argument will have been made that causal determinism is indeed compatible with moral responsibility.[24]

X. APPENDIX: WOLF'S REASON VIEW

A crucial feature of our analysis of guidance control, and thus our theory of moral responsibility, is the switch from an agent-based to a mechanism-based approach. We have motivated this switch by appeal to Frankfurt-type examples, such as "Assassin." In "Assassin," we argued, it is important to distinguish the mechanism that actually leads to the action from the mechanism that *would* have done so (i.e., the mechanism that operates in the alternative scenario). In constructing an account of guidance control, it is indispensable, we have contended, to focus on the properties of the actual-sequence mechanism.

In order further to elaborate and defend this "switch to mechanisms," we wish to discuss an alternative view. Susan Wolf's view – defended in her highly suggestive and very influential book, *Freedom within Reason*, is similar to ours in that she employs the notion of responsiveness to certain reasons ("the ability to choose and act in accordance with the True and the Good"). But the views are different in various respects. Here we shall explain how Wolf's view seems to be a more traditional, "agent-based" view; and we further develop our argument that it is important to switch to a mechanism-based view.

X.1. Wolf's Reason View and the Asymmetry Thesis. Freedom within Reason is a book that, in its own words, is "unabashedly devoted to

[24] Of course, if the Direct Challenge to moral responsibility from causal determinism is valid, then the prima facie argument can be defeated. Thus, the argument that causal determinism is compatible with moral responsibility will not be complete until we turn (in Chapter 6) to an evaluation of the Direct Challenge.

ertheless, it does not seem that the woman is morally responsible for breaking the Steuben egg.

This example illustrates that the WRR test needs to incorporate the idea that an agent's doing otherwise must be *appropriately connected* to the reason to do otherwise. When we consider possible scenarios in which the actual mechanism operates, there is sufficient reason to do otherwise, and the agent does otherwise, we expect that the agent does otherwise *for that reason*.[2] To avoid a lengthy digression, we will not pursue this point much further. However, one final point of clarification is worth noting. Revising the WRR test so that the final clause reads "the agent does otherwise *for that reason*" does not require that the agent actively engage in deliberation or that he consciously consider the sufficient reason in question.[3] From now on, we shall assume that some such revision is implicit in the various formulations of responsiveness. So, for example, under weak reasons-responsiveness

> we hold fixed the operation of the actual kind of mechanism, and we then simply require that there exist *some* possible scenario (or possible world) in which there is a sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise *for that reason*.

The preceding discussion also highlights the fact that, if an agent is to display guidance control, the *actual sequence* must also exhibit the appropriate sort of connection between reasons and action. So, when an agent has guidance control, we assume that he performs the relevant action intentionally (i.e., for a reason). We will take it as implicit in all the accounts of responsiveness that the actual sequence has the appropriate relationship between reasons and subsequent behavior.[4]

[2] The problem here is akin to the worry about wayward causal chains in discussions of reasons as causes of action. For one discussion of this problem, see Donald Davidson, "Freedom to Act," in Ted Honderich, ed., *Essays on Freedom of Action* (London: Routledge and Kegan Paul, 1973), pp. 137–156; reprinted in Davidson, *Essays on Actions and Events* (Oxford: Clarendon Press, 1980).

[3] As Robert Audi has convincingly argued, in order for an agent to act for a reason, *r*, it is not necessary that the person deliberate and formulate *r* as his reason for acting; roughly speaking, it is enough that he would give *r* as the reason for his action, if he were asked for an explanation. See Robert Audi, "Acting for Reasons," *Philosophical Review* 95 (1986), pp. 511–546.

[4] Of course, it is a notorious problem in action theory to specify what this "appropriate relationship" is (in virtue of which an action is intentional). We do not have a specific proposal here.

III. WRR AND THE PROBLEM OF STRANGE PATTERNS

III.1. The Problem. Although WRR is preferable to SRR in allowing for a looser fit between reasons and action, the envisaged fit is, unfortunately, too loose. Consider the recent case in which a person boarded a ferry, and once the boat was underway, produced a saber and proceeded to slay his fellow passengers.[5] Imagine this person to be so disposed that, regardless of how strong the reasons are not to wield his saber, he would still wield the saber (and, upon reflection, approve of the act) in all but *one* possible scenario – a scenario in which he is presented with the reason that he should not kill these people because a passenger is smoking a Gambier pipe in the lower cabin.

Is such an agent properly considered morally responsible for the action? His behavior seems to meet the conditions for weak reasons-responsiveness: there is a possible world in which the actual mechanism (unimpaired practical reason) operates, there is a reason that the saber killer takes to be sufficient to refrain from attacking his fellow passengers, and he does not attack (for the aforementioned reason). Nevertheless, because the saber killer's mechanism of practical reasoning responds to such an unusual reason, we *may* want to say that in fact he is simply manifesting erratic behavior, which, if anything, should count as further evidence of his insanity and consequent lack of responsibility. Although there is a tendency toward this conclusion, it may be that filling in the details of the case in different ways will lead to different conclusions about it; this will indeed be our contention in what follows.

The above example of the saber killer is puzzling largely because we cannot understand the saber killer's motivation and, in particular, why the presence of someone smoking a certain kind of pipe should count as the only reason strong enough to prevent the impending carnage. Normally, when we speak of a mechanism being responsive to reason, we think of a mechanism whose response varies as a function of the *strength* of the reasons presented. In testing responsiveness in different possible worlds, we expect that, as the strength of the reasons is increased, a point will be reached at which the agent, acting on the actual mechanism, will respond differently; moreover, as one moves beyond this threshold, it is assumed that

[5] We owe this sort of example to Ferdinand Schoeman.

increasingly strong reasons will also cause the person to do otherwise.

As it stands, however, a WRR theory does not explicitly require this sort of pattern of responsiveness. Our earlier example (presented in Chapter 2) of the weak-willed person, Jennifer, who will go to the basketball game unless the tickets cost one thousand dollars, is much easier to accept, as a case of moral responsibility, than the saber killer example. We suggest that this is precisely because we *can* understand why a ticket costing a thousand dollars counts as a strong reason not to attend a ball game, but we *cannot* understand why a person smoking a Gambier counts as a strong reason not to dismember ferryboat passengers. But what if it were discovered that, although Jennifer would not attend the basketball game if the tickets cost one thousand dollars, she would go if the tickets cost two thousand dollars? To make the case even more extreme, simply imagine that Jennifer would attend the game in every scenario except the one in which the tickets cost one thousand dollars. At this point, Jennifer's behavior becomes nearly as puzzling as that of the saber killer. Although both of their mechanisms are weakly reasons-responsive, it is not intuitively clear that either is responsible for his or her action. Thus, it appears that weakly reasons-responsive behavior becomes problematic not only when the particular reason to which an agent responds is not easily understood, but also when the general *pattern* of an agent's responses is puzzling.

III.2. Gert and Duggan's Account. Bernard Gert and Timothy Duggan present a responsiveness theory that begins to stress the importance of looking for an *appropriate pattern of response.*[6] They develop a subtle account of free will, free action, and moral responsibility, the finer details of which are not relevant here. Briefly stated, the account requires that, in order for an agent to be morally responsible for some act *A*, he must (1) act intentionally, (2) have the ability to will to do *A*, and (3) not be led to do *A* by coercive incentives.[7] An agent's ability to will to do *A* is analyzed in terms of his ability to believe that there are many and varied, noncoercive (and coercive) incentives for

[6] See Timothy Duggan and Bernard Gert, "Free Will As the Ability to Will," *Noûs* 13 (1979), pp. 197–217; reprinted in John Martin Fischer, ed., *Moral Responsibility* (Ithaca: Cornell University Press, 1986).

[7] Ibid., p. 214.

doing (and not doing) *A*, and its being true that "at least sometimes" in the case of each of several noncoercive incentives (and "almost always" in the case of coercive incentives) the agent aligns his will with his beliefs.[8] What is striking about this account, for the purposes of our discussion, is that it seeks to connect ascriptions of responsibility with an ability to exhibit a certain *pattern* of response to "many and varied incentives." Although this is certainly a step in the right direction, we contend that the kind of criticism raised already against WRR can be extended to the Gert and Duggan view.[9]

To see this point, simply imagine that Jennifer, the ardent basketball fan in the previous example, would *not* go to the game if she believed that the tickets cost $100, $108, $124, . . ., or $137, or $153, or she believed that she had promised to attend her cousin's birthday party, or that it was her turn to do the dinner dishes (or that there were any of a number of coercive incentives not to attend); but she *would* attend if she believed that the tickets cost any of the other prices up to $6,007, or that her uncle invited her to the game, or that a tiny man wearing a Lakers shirt would give her a ticket to the game and a hot dog (or that there were any of a number of coercive incentives to attend). In this example, Jennifer responds – at least sometimes – to many and varied, noncoercive incentives both to attend and not to attend the game (and she responds almost always to

[8] Duggan and Gert's analysis of the ability to will (ibid., p. 210) is as follows:

S has the ability to will to do *X* if and only if

(1) *S* has the ability to believe that there are many and varied coercive incentives for doing a particular act of kind *X*, and almost always, if *S* believed that any of these coercive incentives were present he would will to do that particular act of kind *X*.

(2) *S* has the ability to believe that there are many and varied noncoercive incentives for doing a particular act of kind *X*, and for each of several of these incentives, if *S* believed that it was present he would, at least sometimes, will to do that particular act of kind *X*.

(3) *S* has the ability to believe that there are many and varied coercive incentives for *not* doing a particular act of kind *X*, and almost always, if *S* believed that any of these coercive incentives were present he would will *not* to do that particular act of kind *X*.

(4) *S* has the ability to believe that there are many and varied noncoercive incentives for *not* doing a particular act of kind *X*, and for each of several of these incentives, if *S* believed that it was present he would, at least sometimes, will *not* to do that particular act of kind *X*.

[9] David Shatz makes this point, along with raising a number of other interesting criticisms of Duggan and Gert's account, in his article, "Compatibilism, Values, and 'Could Have Done Otherwise,'" *Philosophical Topics* 16 (1988), pp. 151–200.

coercive incentives); but still her pattern of response is so strange that it raises the question of whether Jennifer, in exhibiting this pattern of counterfactual response, can be held morally responsible at all.

So, whereas Gert and Duggan require more than mere WRR, their requirements are still too weak. The problem seems to be that their more stringent requirements still do not impose enough "structure." The purely "quantitative" vocabulary they employ – "many and varied," "at least sometimes," "almost always," and so forth – still leaves room for weird patterns that would seem to rule out moral responsibility.

III.3. The Challenge. Where does this leave the theorist who hopes to connect responsibility and responsiveness? As we have seen, strong reasons-responsiveness is too restrictive to serve as a part of a set of necessary and sufficient conditions for morally responsible behavior.[10] For example, this condition is so strong that it withholds responsibility ascriptions from agents who merely act in a weak-willed manner. Loosening the requirement for responsibility to weak reasons-responsiveness (or to Gert and Duggan's more nuanced requirement) properly ascribes responsibility to weak-willed agents, but the condition then becomes so loose that it also ascribes responsibility to agents who act on mechanisms that respond only in unusual or incoherent ways. The challenge apparently facing a responsiveness theorist, then, is to find something of a middle ground between SRR and WRR, in which there is sufficient structure in the profiles of responsiveness to reasons relevant to moral responsibility.[11]

[10] As pointed out in Chapter 2, "tracking value" a la Nozick is also too strong for moral responsibility.

[11] A detailed discussion of the nature of reasons for action is beyond the scope of this work. Indeed, we hope to present an account of moral responsibility that is consistent with the broadest possible array of views about the nature of reasons for action. Of course, philosophers differ about what reasons for action are (desires, beliefs, desire–belief pairs, states or conditions of the external world, and so forth). Also, apart from this "ontological" dispute, they disagree about the particular conditions in which it is true to say that something constitutes a reason for action. We shall not enter into such disputes here. We want a theory of responsibility that can fit with the widest possible selection of plausible views about reasons for action; it would be undesirable if one's theory of moral responsibility depended on a contentious theory of reasons for action.

IV. MODERATE REASONS-RESPONSIVENESS

IV.1. Receptivity. Recall that if an actually operative mechanism is strongly reasons-responsive, it meets three conditions: (1) it is strongly *receptive* to reasons – that is, the agent would recognize what reasons there are, given that the actual kind of mechanism operates; (2) it is strongly *reactive* to reasons – that is, the agent would choose in accord with the reasons recognized, given that the actual kind of mechanism operates; and (3) it produces actions that are in accord with choice – that is, the agent, when he is acting on the mechanism, would act in accord with his choice. On this schema, a mechanism moves from being strongly reasons-responsive to being weakly reasons-responsive (or not responsive at all) as a result of a "deficiency" in any of these three areas.

For our purposes in the remainder of this chapter, it is convenient to treat the final two categories as one. So, by "reactivity to reason," we shall mean the capacity to *translate* reasons into choices (and then subsequent behavior). Of course, by "receptivity to reason," we shall mean the capacity to recognize the reasons that exist.

We contend that the reactivity to reasons and receptivity to reasons that constitute the responsiveness relevant to moral responsibility are crucially *asymmetric*. Whereas a very weak sort of reactivity is all that is required, a *stronger* sort of receptivity to reasons is necessary for this kind of responsiveness. To help to motivate the asymmetry claim, consider, first, the case of Brown, a weak-willed individual with a strong craving for the nonaddictive drug "Plezu." (By saying that the drug is "nonaddictive," we here mean that it does *not* issue in *irresistible* urges to take it.) Plezu directly stimulates the pleasure centers in one's brain, causing euphoria. If used infrequently, it produces no harmful side effects. The main difficulty with Plezu is that it is so powerful that its effects last for hours and, during this time, it renders the user unable to do anything except recline on the sofa and enjoy himself. As a result, frequent use of Plezu typically results in loss of job, family, and self-respect.

Brown, unfortunately, is so fond of Plezu and so lacking in self-discipline that, although he recognizes that there are strong reasons *not* to take the drug every morning, he typically ends up passing his day on the couch. In fact, let us say that the only scenario in which Brown would not take Plezu is one in which he is told that injecting the drug once more would have an extremely grave con-

sequence – death (a side effect of Plezu that threatens only longtime users).

In this example, Brown acts on a mechanism that is only weakly *reactive* to reasons. Brown is weak-willed, but we believe that he is, nevertheless, morally responsible for his action of taking the drug. The fact that he refrains from injecting Plezu in at least one instance (holding fixed the operation of his actual-sequence mechanism) is important evidence that Brown has the sort of control associated with moral responsibility for actions (guidance control). The weak reactivity of Brown's actual-sequence mechanism is a reflection of the fact that his urges to take Plezu are *not* irresistible. We claim that it is plausible that weak reactivity to reasons is all the reactivity required for guidance control (and moral responsibility); here this is only a plausibility claim, which we shall defend in the following section.

In contrast, weak receptivity to reasons is *not* all the receptivity required for moral responsibility. To see this, consider a modified version of the previous example, in which Brown is now acting on a mechanism that is reactive to reasons but *only weakly receptive* to them. Imagine, for example, that Brown would agree that, if the Plezu cost one thousand dollars per injection, this would be a sufficient reason for him not to take the drug. (Let us say that, in this circumstance, Brown would not take the drug.) Brown appears to show that in at least one instance he *recognizes* a sufficient reason not to take the drug; hence, given that he would react appropriately to this reasons-recognition, someone might argue that he is responsible for his action.

But what if we then discovered that Brown (when acting on the same mechanism) would *not* recognize that prices of two, or three, or four thousand dollars also counted as sufficient reasons not to take the drug? If we further discovered that, regardless of how strong the other reasons were not to take the drug, Brown would *only* recognize the thousand dollar price to be a sufficient deterrent, then, presumably, we would wonder whether Brown is indeed morally responsible.

This example suggests that being weakly receptive to reasons is not sufficient to show that a person is acting on a mechanism that is receptive to reasons in the sense required for moral responsibility. In judging a mechanism's receptivity, we are not only concerned to see that a person acting on that mechanism recognizes a sufficient reason in one instance; we also want to see that the person exhibits an

appropriate *pattern* of reasons-recognition. In other words, we want to know if (when acting on the actual mechanism) he recognizes how reasons fit together, sees why one reason is stronger than another, and understands how the acceptance of one reason as sufficient implies that a stronger reason must also be sufficient.

In order for an agent to be morally responsible for an action, then, the actual mechanism that issues in his action must be at least "*regularly*" receptive to reasons.[12] A person who acts on a regularly receptive mechanism must exhibit a certain sort of pattern of reasons-recognition.[13] More specifically, our suggestion is that it is a defining characteristic of regular reasons-receptivity that it involves an *understandable pattern* of (actual and hypothetical) reasons-receptivity.[14]

On our approach, it is as if a "third party" (the one assessing the moral responsibility of the relevant agent) conducts an "imaginary interview" with the agent. In this interview, he asks about various actual and hypothetical scenarios, and elicits views from the agent as to what would constitute sufficient reasons. Even if a person claimed that, given his actual values (or preferences), only one reason counts as sufficient, the pattern of his actual mechanism's receptiveness could still be tested by asking him which reasons would count as sufficient, given *another* set of values (or preferences). The third party then employs the information from the interview, together with background information, to seek to understand the pattern in the set of reasons-recognitions. What is required is that the configuration of

12 We use the term "regularly" here not in the sense of "normally" or "customarily," but only in the weaker sense that implies a degree of orderliness and regularity.

13 Of course, the regularity need not be absolute; the mechanism must simply evince some suitable degree of regularity. Everyone makes some mistakes, and it is a matter of judgment precisely how much regularity is appropriate to require.

14 For this idea, see Mark Ravizza, "Moral Responsibility and Control: An Actual-Sequence Approach" (Ph.D. diss., Yale University, 1992); also, see Shatz, "Compatibilism, Values, and 'Could Have Done Otherwise,'" p. 177. There is also useful discussion of this sort of approach in Paul Benson, "Responsibility, Reasons-Responsiveness, and Self-Worth," unpublished manuscript. Of course, we recognize that the notion of an "understandable pattern" is still quite vague. Note that "understandable pattern" is intended to imply something more than *mere consistency;* rather, the agent's recognitions of reasons must fit together in a more robust sense. The *contents* of the responses must *interact in a substantive way* – a way that is, admittedly, difficult to specify generally – to produce a pattern that is understandable to a suitably placed third party.

answers in the imaginary interview can (together with background information) give rise to an understandable pattern, from the perspective of the third party (the person judging whether the agent is morally responsible).

It is a constraint on the third party, as we envisage him, that he requires that certain "objective" conditions be satisfied, in assessing the patterns in an agent's reasons-receptivity. Relative to an agent's preferences, values, and beliefs, reasons are graded in terms of their strength. An agent's reasons-receptivity must exhibit a suitable correspondence to the objective (relative to a given set of preferences, values, and beliefs) grading of the strength of reasons, in order for the pattern in the agent's reasons-receptivity to be understandable.[15] For example, if a ticket's costing a thousand dollars is a reason not to go to the game, surely (barring unusual circumstances) a ticket's costing two thousand dollars should be a reason not to go to the game, and so forth.

When "reasons" do not connect and relate to one another in appropriate ways, they do not generate a minimally comprehensible pattern. Recall the example of the saber killer, as originally presented. Remember that the example is puzzling because we cannot understand the saber killer's motivation and, in particular, why the presence of someone smoking a particular pipe should count as the only "reason" strong enough to prevent him from killing passengers with his saber.

But if we *could* understand why this was the only reason sufficient for the saber killer to do otherwise, our intuitions about his responsibility would quickly change. Imagine that we discover that, as in the last chapter of an intricate tale of mystery and espionage, the saber killer actually was a secret service agent who had an obscure, but understandable, reason for terminating certain passengers on the ferry. His mission, which involves the fate of humanity, is of such importance that the only thing that could possibly count as a good reason for changing his plan is to see a fellow secret service agent smoking a distinctive Gambier pipe, because this serves as a sign that the killing of these passengers has been rendered unnecessary thanks to the activity of other agents working in Czechoslovakia. In this case, people would no longer feel uncertain about ascribing

[15] The "objective" grading of the strength of reasons (relative to the given preferences, values, and beliefs of the agent) is simply given by the wide reflective equilibrium of the community. (We do not mean anything more by "objective.")

responsibility to the saber killer. In fact, we probably would be inclined to praise him for his action.

Regular reasons-receptivity, then, is reasons-receptivity that gives rise to a minimally comprehensible pattern, judged from some perspective that takes into account subjective features of the agent (i.e., the agent's preferences, values, and beliefs) but is also *not simply* the agent's point of view. A comprehensible pattern of reasons-recognitions may, however, be utterly divorced from reality. Thus, we claim that the agent's answers in the imaginary interview must *also* be at least minimally "grounded in reality." That is, regular receptivity to reasons, in the sense that is required for guidance control and moral responsibility, requires at least that the agent not be substantially deluded about the nature of reality. Regular receptivity to reasons, then, requires an understandable pattern of reasons-recognition, minimally grounded in reality.

IV.3. Reactivity. In the previous section we developed a notion of reasons-receptivity that is more robust than mere weak receptivity, and we contended that this notion of regular receptivity is required for moral responsibility. We suggested that, in contrast, mere weak reactivity to reasons is all the reactivity required for moral responsibility. Whence the asymmetry?

Suppose that, in the example developed in the previous section, Brown said this: "It is unfair to hold me morally responsible for taking Plezu. After all, although I am regularly receptive to reasons, I am only weakly reactive to reasons. Thus, whereas I would have responded to a very different incentive for doing otherwise, the mechanism on which I acted did not – and *could not* – have responded to the *actual* incentive to do otherwise. Given this, it is unfair to hold me morally responsible."

We believe that a cogent reply to Brown is available. This reply is based on the fundamental intuition that "reactivity is all of a piece." That is, we believe that if an agent's mechanism reacts to *some* incentive to (say) do other than he actually does, this shows that the mechanism *can* react to *any* incentive to do otherwise. Our contention, then, is that a mechanism's reacting differently to a sufficient reason to do otherwise in some other possible world shows that the same kind of mechanism can react differently to the *actual* reason to do otherwise. This general capacity of the agent's actual-sequence mechanism – and *not* the agent's power to do otherwise – is what helps to ground moral responsibility. (As we claimed in Chapter 2,

an important lesson of the Frankfurt-type examples is that we should shift our focus from features of the agent to properties of the actual-sequence mechanism from which he acts.) So it is plausible to reply to Brown that his mechanism of practical reasoning (the mechanism that actually produced his behavior) could in fact have reacted to his actual reason not to take Plezu.

The picture here is of one kind of mechanism with different "inputs." Further, the idea is that reactivity is all of a piece in the sense that the mechanism can react to all incentives, if it can react to one. How might this have turned out not to be true? Imagine that the agent somehow gets considerably more energy or focus if he is presented with a *strong* reason to do otherwise, and it is only in virtue of these factors that he succesfully reacts to the reason. There certainly can be cases like this, but it is natural to say that, when the agent acquires significantly more "energy or focus," this gives rise to a *different mechanism* from the actual mechanism. Our point is that, *holding fixed the actual kind of mechanism,* reactivity is all of a piece: if the mechanism can react to any reason to do otherwise, it can react to all such reasons.[16]

It should be evident that there can be considerably more idiosyncrasy in the reactivity component of the mechanism that leads to action than in the receptivity component. Regular receptivity to reasons is required for moral responsibility; thus, there must be an understandable pattern in the profile of an agent's recognitions of reasons (holding the actual mechanism fixed). But the situation is different with respect to the reactivity component. As we pointed out, an agent may be morally responsible even though his mechanism is only weakly reactive to reasons (reacting differently in only a very small set of worlds). Further, a responsible agent may exhibit a bizarre pattern of reactivity. (Some agents may even deliberately choose to exhibit such a pattern.) Even if the pattern of reactivity is bizarre, we contend that if the agent's mechanism reacts to a sufficient reason to do otherwise, it *can* react to any reason to do otherwise. This general power of the mechanism explains why even a highly idiosyncratic pattern of reactivity is consistent with moral responsibility.

[16] One might wonder how our account applies to those situations in which it is absolutely clear what the agent should do, and thus there doesn't appear to be any incentive to do otherwise. Note that there is *always* at least *some* reason to do otherwise, even if it is not a good or "sufficient" reason to do otherwise. Of course, an agent need not explicitly consider or consciously focus on the reasons that are, nevertheless, available. And even in a context in which there are no "good" reasons for the agent to do otherwise, there are still reasons for the agent to do otherwise. (Consider, for example, that anyone can *wonder* whether his actual sort of mechanism could respond differently; thus, a reason to do otherwise would be to prove that the mechanism is indeed responsive. For further discussion, see John Martin Fischer and Mark Ravizza, "When the Will Is Free," in James E. Tomberlin, ed., *Philosophical Perspectives VI: Ethics* (Atascadero, Calif.: Ridgeview, 1992), pp. 423–451; and John Martin Fischer, *The Metaphysics of Free Will: An Essay on Control,* Aristotelian Society Monograph Series, vol. 14 (Cambridge, Mass.: Blackwell Publishers, 1994), pp. 47–58.

The asymmetry between reactivity and receptivity can now be stated crisply. In the case of receptivity to reasons, the agent (holding fixed the relevant mechanism) must exhibit an understandable pattern of reasons-recognition, in order to render it plausible that his mechanism has the "cognitive power" to recognize the actual incentive to do otherwise. In the case of reactivity to reasons, the agent (when acting from the relevant mechanism) must simply display *some* reactivity, in order to render it plausible that his mechanism has the "executive power" to react to the actual incentive to do otherwise. In both cases the pertinent power is a general capacity of the agent's mechanism, rather than a particular ability of the agent (i.e., the agent's possession of alternative possibilities – the freedom to choose and do otherwise).

To illustrate the asymmetry claim, imagine that Brown abstains from taking Plezu only when the injections cost one thousand dollars. In other cases, even if the injections are more expensive, he still takes the drug. Suppose, further, that Brown announces that he recognizes that the more expensive prices also constitute sufficient reasons not to take the drug, but adds that he just wants to act on a whim and abstain from Plezu only when it costs one thousand dollars.[17] Given that Brown's mechanism is suitably receptive to reasons, we believe that he would be morally responsible for his behavior here, even though he is acting from a merely weakly reactive mechanism. In contrast, imagine that Brown were sincerely to say that he recognizes that a one thousand dollar price is a sufficient reason not to take the drug because it is too expensive, but then added that he does not understand why the higher prices are also sufficient reasons to abstain. Here there is a strong intuition that (holding fixed the relevant mechanism) Brown does not display a

[17] This is, of course, simply one way of exhibiting weak reactivity. There are other agents who do not deliberately adopt weak reactivity.

sufficient understanding of how reasons work to hold him responsible for the action.

In this (and the previous) section, we have employed the distinction between reasons-recognition and reasons-reactivity to generate an account of reasons-responsiveness that is richer than WRR. This sort of responsiveness requires only *weak* reactivity to reasons, but *regular* receptivity to reasons. Regular receptivity to reasons requires a pattern of actual and hypothetical recognition of reasons that is understandable and minimally grounded in reality. We have moved some distance away from the more schematic WRR, without demanding anything so strong as SRR. But we still need to refine the account a bit more, in order to fill out our understanding of moderate reasons-responsiveness. Having done so, we shall return to the example of the saber killer.

IV.3. Smart Animals, Children, and Psychopaths. A further refinement to this account is recommended by considering borderline cases involving creatures like intelligent animals, very young children, and psychopaths. All of these creatures (arguably) exhibit a certain pattern of responsiveness to reason; nevertheless, none of them is ordinarily judged to be morally responsible. This is, we suggest, because such creatures are not *moral* agents. Although they may act on mechanisms that respond to instrumental or prudential reasons, they are not appropriately responsive to moral demands. We suggest that these individuals are not moral agents (and not properly held morally responsible for their behavior) because they are without any understanding and appreciation of moral reasons.[18]

Of course, it is a vexed and highly contentious matter how to specify "moral reasons." We shall operate with a fairly simple way of differentiating moral from "prudential" reasons. On this approach, prudential reasons concern an agent's long-term self-interest, whereas moral reasons issue from some sort of (suitable) balancing of one's own interests against the interests and rights of others. It is not here assumed that moral reasons are "correct moral reasons" (on any account of correctness of moral reasons). But what count as

[18] Peter Arenella stresses the importance of being responsive to moral reasons in raising an analogous criticism of rational choice theorists; see Peter Arenella, "Character, Choice, and Moral Agency: The Relevance of Character to Our Moral Culpability Judgments," *Social Philosophy & Policy* 7 (1990), pp. 59–83. See, also, R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, Mass.: Harvard University Press, 1994), p. 189.

moral reasons, and what are at least in the "ball park" as contenders for being correct, are given by the considered judgments (in wide reflective equilibrium) of the relevant community.

A natural way, then, to handle individuals such as intelligent animals, young children, and psychopaths is to refine the responsiveness theory so that it more explicitly requires not merely responsiveness to prudential reasons, but to (at least some) moral reasons as well. More specifically, we shall understand our responsiveness theory to require that an agent be regularly receptive to reasons, at least some of which are moral reasons. On this view, responsibility requires that (given the actual-sequence mechanism) the agent recognize an understandable pattern of moral reasons; that is, just as the recognition of nonmoral reasons must have a suitable *structure*, so must the recognition of moral reasons.[19]

As before, the requirements for a regularly receptive pattern of recognition are fairly schematic. For instance, such a pattern of receptiveness must include a recognition that certain moral demands are stronger than others, that in some instances the rights of others outweigh one's own prudential interests, and so forth. It is not necessary that the agent (when acting on the actual mechanism) recognize each and every moral reason. Nor does the theory depend on the assumption that there is some objective moral standard against which every agent is judged. Finally, it is not required that responsible agents, even within a given moral community, exhibit exactly the same pattern of recognizing, weighing, and ranking moral and prudential reasons. A regularly receptive pattern should evince that the agent (when acting on the actual mechanism) recognizes both that other persons in the community have claims and that these claims give rise to reasons for action. That is, the pattern in question must show that the agent (when acting on the actual mechanism) recognizes that other persons' claims give rise to moral reasons *that apply to him.* Without such a minimal receptiveness to moral reasons, agents would fail to be moral agents at all, and consequently would not be appropriate candidates for the reactive attitudes.

Rather than trying to characterize an appropriate pattern of receptiveness in greater detail, it is perhaps more illuminating to outline the boundaries of such receptiveness by citing several paradigm

[19] Again, it does not seem fruitful – or necessary – to specify precise numerical (or other) requirements for receptivity to moral reasons. The theory's vagueness is perhaps an inevitable reflection of the nature of the phenomena it purports to analyze.