SELF-DEFEAT, PUBLICITY, AND INCOHERENCE:

THREE CRITERIA FOR CONSEQUENTIALIST THEORIES

by

J. Benjamin Eggleston, Jr.

B.A., Washington and Lee University, 1994

M.Sc., University of Edinburgh, 1995

M.A., University of Pittsburgh, 1999

M.A., University of Pittsburgh, 2001

Submitted to the Graduate Faculty of the
Faculty of Arts and Sciences in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy

University of Pittsburgh

2002

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

J. Benjamin Eggleston, Jr.

It was defended on

December 18, 2001

and approved by

David DeJong, Professor of Economics

Edward McClennen, Centennial Professor of Philosophy

Nicholas Rescher, University Professor of Philosophy

Michael Thompson, Associate Professor of Philosophy

Dissertation Director: David Gauthier, Distinguished Service Professor of Philosophy

SELF-DEFEAT, PUBLICITY, AND INCOHERENCE:
THREE CRITERIA FOR CONSEQUENTIALIST THEORIES

J. Benjamin Eggleston, Jr., Ph.D.

University of Pittsburgh, 2002

This dissertation identifies and assesses three criteria that are often used to evaluate consequentialist theories of morality and rationality. After introducing a distinction between straightforwardly maximizing consequentialist theories (such as act consequentialism and egoism) and indirectly maximizing consequentialist theories (such as rule consequentialism, rule egoism, Gauthier's theory of constrained maximization, and McClennen's theory of resolute choice), it addresses criteria associated with the concepts of self-defeat, the publicity condition, and incoherence. It argues that (1) the thesis that the self-defeat of a normative theory is a good reason for rejecting it has several surprising and intolerable implications, (2) the publicity condition is an unreasonably demanding requirement to impose on normative theories, and (3) unlike self-defeat and publicity, the issue of incoherence is crucial to the viability of a normative theory; consequently, the incoherence of indirectly maximizing theories renders them unacceptable as accounts of moral or rational action. Although each of these conclusions is of independent interest, they are of further interest when considered together. For cumulatively they constitute a vigorous defense of straightforwardly maximizing theories and a sharp indictment of their indirectly maximizing rivals. As a result, the dissertation has direct implications for debates in both normative ethics and rational-choice theory.

TABLE OF CONTENTS

LIST OF FIGURES

I

Theories and Standards

1 Normative theories

1.1 What reasons for accepting and rejecting normative theories are good ones? Should it matter, for example, if a theory is *self-defeating*, as egoism is often said to be? Should it count as a strike against a theory if it violates the *publicity* condition, as act consequentialism is sometimes said to do? Should rule consequentialism be rejected if it is *incoherent* in the way in which some critics say that it is? This dissertation is an attempt to answer these questions.

This dissertation is, then, about things of two kinds. First, it is about what I call *normative theories*: theories that provide rules or prescriptions that tell agents how they ought to act. Some familiar examples of such theories are the many variants of egoism, consequentialism, Kantianism, and contractarianism. Second, it is about the considerations that we should regard as good reasons for accepting and rejecting normative theories or, as I shall refer to them, the *criteria* or *standards* that it is appropriate to use in the evaluation of normative theories.

As I said, this dissertation is an attempt to ascertain how seriously we ought to take certain standards—those having to do with self-defeat, the publicity condition, and a certain kind of incoherence. As one might suspect, the motivation for such an inquiry is to improve our reckoning of the merits and demerits of the theories that are often criticized in these terms. Given this underlying purpose, it would be natural to wonder at the outset what the point could be of dwelling on the standards. Wouldn't it make more sense to focus on the theories themselves?

Three considerations suggest not. First, focusing on the theories themselves would quickly lead to an inquiry of the present kind, anyway, since the theories themselves are, as we shall see, often—indeed, recurrently—criticized in terms of the standards we shall be examining. Second, the standards themselves need to be scrutinized directly: they are so frequently invoked at the *end* of trains of reasoning, as the deepest and most dispositive considerations that can be brought to bear on an issue, that their validity or rational warrant is often uncritically assumed. By taking the standards as our logical starting point, we should be able to avoid treating them as the last word. Third, as we shall see shortly, the standards we shall be examining are used to evaluate many different theories, and even theories of different kinds. Focusing on the standards will enable us to survey their operation in regard to many different theories and will afford us a richer understanding of them than can be accessed by way of a direct inquiry into the merits and demerits of the theories themselves.

1.2 Since we shall be examining standards for evaluating normative theories, some account of the notion of a normative theory is in order. I said above that normative theories are theories that provide rules or prescriptions that tell agents how they ought to act. One feature of such theories that must be made explicit is that not all normative theories operate in the same normative domain: some deal with morality, some with instrumental rationality, some with etiquette, and so on. Act consequentialism is, of course, a theory of morality, while egoism shall be construed here as a theory of instrumental rationality.

A second feature of such theories is that they do not necessarily purport to tell agents what they ought to do, all things considered. For our purposes, a theory counts as a normative theory as long as it purports to tell agents what they ought to do in the domain in which it operates. For example, egoism counts as a theory of instrumental rationality in virtue of purporting to tell agents what it is rational and irrational to do, even if it does not claim that it is never the case that one ought—in some ultimate sense of ‘ought’—to act contrary to its prescriptions. Similarly, act consequentialism counts as a normative theory in virtue of purporting to tell agents what it is right and wrong to do, even if it does not also say that moral considerations override all others and thereby purport to tell agents what genuine, all-encompassing, rationality would bid them do. So a normative theory needn’t, in effect, assert the primacy of its domain over other normative domains.

The point of this stipulation is to be as inclusive as possible in our designation of theories to be considered. For although there are many moral philosophers who regard moral considerations, when properly sorted and weighed, as including and hence trumping all others,[1]

[1] Crisp writes that for Mill, “there are no principles, moral or otherwise, to compete with utilitarianism: it is not just *moral* rationality, but rationality *tout court*, that requires the maximization of utility by each person” (p. 81), meaning that “prudence has no weight against the demands of the greatest happiness principle” (p. 149). And Hare writes that “moral principles are . . . superior to or more authoritative than any other kind of principle” (1981, p. 169).

there are also some who shy away from such claims.[2] The theories advocated by the former undoubtedly raise interesting issues that the theories advocated by the latter do not, but none of those is among the issues with which we shall be concerned.[3]

A third feature of normative theories is that they do not necessarily provide an account, or an analysis, of the normative terms to which they appeal. The upshot of this feature is most easily seen in the context of moral theories. For there is a certain distinction frequently drawn in any of several ways: between ethics and meta-ethics, or between normative and analytical ethics, or between first- and second-order theorizing.[4] On the conception of moral theories operative here, they needn't be meta-ethical theories: they do not need to have anything to say about the meanings of moral words, or what is often called the "status" of moral judgments. And the same, *mutatis mutandis*, goes for normative theories of other kinds.

1.3 It follows from the foregoing that normative theories focus on *acts*. Norms apply, of course, to other things, such as motives, dispositions, rules, and institutions; and a normative theory may well offer assessments of some or all of these other things. Indeed, depending on the connections a normative theory makes between acts and these other objects of normative assessment, a normative theory's assessments of these other things may be almost entirely determined by—or even determinative of—its assessments of acts. But for our purposes something may be a normative theory even if it neglects these other matters entirely, as long as it is (explicitly) action-guiding.

Philosophers have long been concerned with action-guiding theories. This is especially true of moral philosophers, many of whom take themselves to be engaged in the task of developing normative theories that take, of course, the point of view of morality. Mill, for example, writes that "It is the business of ethics to tell us what are our duties" (1861, p. 219 [ch. II, par. 19]), and Rawls voices this sentiment more explicitly in his claim that

[2] Sidgwick famously regards the superiority of the moral to the prudential as "the profoundest problem of Ethics" (p. 386, n. 4) and laments that practical reason is beset by "an ultimate and fundamental contradiction" between the demands of prudence and those of morality (p. 508).

[3] It has been claimed that a moral theory as understood here would be an impoverished one. Scanlon, for example, claims that "A satisfactory moral philosophy will not leave concern with morality as a simple special preference, like a fetish or a special taste, which some people just happen to have. It must make it understandable why moral reasons are ones that people can take seriously, and why they strike those who are moved by them as reasons of a special stringency and inescapability" (p. 106). But it is possible both to acknowledge the importance of the explanations that Scanlon says a satisfactory moral philosophy must provide and to point out that the standards that we shall be examining do not require, for their application, that theories offer such explanations.

[4] For brief and clear accounts of such distinctions, see Hare (1963, p. 97; 1971a, p. 127; and 1997, p. 45) and Mackie (1977, p. 9).

> in ethics we are attempting to find reasonable principles which, when we are given a proposed line of conduct and the situation in which it is to be carried out and the relevant interests which it affects, will enable us to determine whether or not we ought to carry it out and hold it to be just and right. (1951, p. 2)

This position is advanced yet more strongly in the following remark by Narveson:

> Let us begin by recalling the primary function of ethical principles: to tell us what to do, i.e., to guide action. Whatever else an ethical principle is supposed to do, it must do that, otherwise it could not (logically) be an ethical principle at all. (1967, p. 105)

According to Singer, developing normative theories has been "[t]he historic task of moral philosophy" (1974, p. 515), and Kagan writes that the "received view" is that "normative ethics . . . is concerned with determining which actions are right, which wrong," among other things (1992, p. 223). Baier explains that "action has, for some time now, occupied the limelight in ethics because, although we have a measure of control over [such things as "behavioral dispositions, character traits, habits or feelings"], we have the most direct control over [acts]: they are things we can do or perform at will" (1995, pp. 282–283).

To be sure, some moral philosophers object to this emphasis on individual acts. Baier is careful to note that his approach "does not imply the priority of action over personality, nor the other way around" (1996, p. 266). And Adams, citing Narveson's remark, laments the prevalence of "the assumption that 'What should I (try to) do?' is *the* ethical question, and that we are engaged in substantive *ethical* thinking only insofar as we are considering *action*-guiding principles" (p. 474). According to Adams, the prevalence of this assumption prevents us from seeing the point of genuinely ethical questions such as "What motives should I have?" by causing us to read such questions as asking "What motives should I try to develop and maintain in myself?" (p. 474).

Moreover, some philosophers object to the thought that moral theories are action-guiding in the sense in which that term has been used: as referring to the issuing of specific prescriptions for action. It has been claimed, for example, that

> moral rules count as a starting point, but their role is not to "tell us the right thing to do." This alternative conception of moral rules fosters creativity and imagination in addressing moral problems, rather than rigidity and frustration. (Morgan, p. 105)

On this alternative conception of moral rules, they may serve to communicate the wisdom that accumulates with experience, thereby bringing into view features of situations that many people would find salient, but they do not amount to outright prescriptions.

Finally, the history of moral philosophy includes a long and venerable strand that denies the very possibility of constructing a sound moral theory that is action-guiding in the way presupposed here. This tradition begins with Aristotle's famous remark that once the relevant moral considerations have been brought into view, "The decision rests with perception" (p. 1109b, l. 23; cf. p. 1126b, l. 4), and is echoed in Ross's claim that once an agent's prima facie duties have been identified, only judgment can reveal which is the strongest (p. 19). And Williams warns that

> There cannot be . . . an ethical theory, in the sense of a philosophical structure which, together with some degree of empirical fact, will yield a decision procedure for moral reasoning. (1981, pp. ix–x)

In his *Ethics and the Limits of Philosophy* (1985), Williams deliberately takes as his focus a question—the question of how one should live—that, he says, "stands at a distance from any actual and particular occasion of considering what to do" (p. 19).

These challenges to the idea of an action-guiding theory raise difficult and deep questions that, instead of trying to answer, I shall have to set aside without further consideration. For no matter how they ought to be answered, we shall see that important issues are raised, and important lessons may be learned, through a consideration of standards that apply to theories simply in virtue of their focusing on individual acts and purporting to offer prescriptions that amount to conclusive verdicts, not just ingredients for the exercise of perception or judgment or as stimulants of creativity and imagination.

2 Morality and rationality

2.1 I said above that a normative theory may operate in any of several different domains: morality, instrumental rationality, etiquette, and others. Indeed a normative theory could equally well take more-specialized points of view, such as respect for family, long-term financial success, or cleverness. So the kinds of normative theories that are possible are as unlimited as is our vocabulary of evaluative terms. In this dissertation, I shall be concerned with normative theories of just two kinds: theories of morality, or moral theories, and theories of instrumental rationality (though I shall typically omit the qualifier ‘instrumental’).

What are such theories about? Acting morally characteristically involves acting on some concern for others above and beyond the concern for others that may already be incorporated into the agent’s own interests and ends. Mill expresses this thought in his claim that the feeling of justice “derives its morality from” the fact that it includes “all persons, by the human capacity of enlarged sympathy, and the human conception of intelligent self-interest” (1861, p. 250 [ch. V, par. 23]). Similarly, Kupperman claims that “the more extreme forms of egoism . . . arguably . . . are not moral views, but rather the refusal to enter into moral reflection or judgment at all” (1981, p. 308). Finally, Crisp asks, “what is morality if it does not involve some concern for others?” (p. 79). This, then, is the perspective from which prescriptions are typically issued by theories of morality.

What about theories of rationality? Unlike theories of morality, they are essentially concerned only with the agent’s *own* interests and ends; such theories may regard, as rational, an act that any plausible moral theory would regard as immoral. Now recall also that theories of rationality are, as I said, theories of *instrumental* rationality. This means that when a theory of rationality declares certain acts to be rational in contrast to others that are irrational, it does not thereby imply just that those acts are supported by reasons in some general sense, and certainly not that they supported by reasons in some ultimate sense. Rather, a theory of rationality declares certain acts to be rational on the basis of those acts’ being related in some way to the advancement of the agent’s self-interest, or to the agent’s life’s going as well as possible. Such is the nature of instrumental rationality, whatever else may be said about rationality *tout court*.[5]

These are the things that we shall take moral theories and theories of rationality to be about. But how can one tell whether a normative theory is a moral theory, a theory of rationality, or a normative theory of some other kind? It will be sufficient for our purposes to plan on finding out just by looking at the predicates or terms it assigns to certain acts. If it calls some acts ‘right’ and some ‘wrong’, then it’s a moral theory. If it calls some acts ‘rational’ and others ‘irrational’, then it’s a theory of rationality. If it calls some acts ‘clever’ and others ‘not clever’, then it’s a theory of cleverness. From this simple rule it follows that the *substance* of the principles by

[5] One aspect of instrumental rationality not explored in this dissertation is what constitutes a person’s self-interest, or what is involved in a person’s life’s going as well as possible. This question is discussed by many authors, including Griffin (1986), Rescher (1993, esp. ch. 3 [pp. 45–62]), and Sumner.

which a theory makes these distinctions does not determine the type of theory it is: if a theory calls some acts ‘clever’ and others ‘not clever’, then we shall regard it as a theory of cleverness, even if the principles on which it makes this distinction don’t seem to have anything to do with cleverness. Such a theory would probably fail to be a good theory of cleverness, but it would not fail to be a theory of cleverness. What a theory is a theory *of* is whatever it purports to be a theory of.

2.2 Many writers who concentrate on moral theories and theories of rationality, as I do, are thereby led to discuss whether and to what extent the dictates of morality can be reconciled with those of rationality. Ever since Glaucon and Adeimantus challenged Socrates to show that being just really is good for the just person, and not just good for other people with whom that person may come into contact, moral philosophers have been interested in showing a coincidence between the dictates of their favored moral theory and the dictates of (instrumental) rationality. Almost without exception, modern moral philosophers have regretfully found that “it appears improbable that this coincidence is complete and universal” (Sidgwick, p. 175), and even those who show the empirical tendency of morality and self-interest to coincide admit the theoretical divergence (Rescher 1975, p. 69). Indeed some have held that morality *inherently* requires some divergence from self-interest; Sterba goes so far as to say that a principle is not a moral principle if “it never requires a person to sacrifice her overall interest for the sake of others” (p. 113).

Now I said earlier (in subsection 1.2) that a theory needn’t, in order to be a normative theory, assert the primacy of its domain: that a moral theory needn’t declare moral considerations to override all others, and that a theory of rationality needn’t purport to provide all-things-considered ‘ought’ judgments, either. Here I add the related point that such questions not only needn’t be addressed by the theories to be concerned, but also shall not be addressed in this dissertation. For although such questions are of lasting interest, they are not raised by the standards that we shall be examining.

3 Consequentialist theories: straightforwardly and indirectly maximizing

3.1 In this dissertation, we shall be concerned with only a subset, though a very important one, of moral theories and theories of rationality: consequentialist ones.

A consequentialist theory is, roughly, one according to which whether acts are right or wrong, or rational or irrational, or otherwise permissible or impermissible, depends entirely on the consequences that they bring about.[6] The canonical consequentialist moral theory is, of course, classical utilitarianism. As systematically articulated by Bentham and brought to relative philosophical maturity by Sidgwick, this theory holds that an act is right if it maximizes overall happiness. By abstracting from utilitarianism, we can formulate a theory with which the standards that we shall be examining are concerned:

> An act is right if and only if its consequences are at least as good as those of any other act that the agent could have performed instead.

I shall call this theory *act consequentialism*.[7] Admittedly, questions may be raised about the propriety of this label; it might be thought that a better label would be 'maximizing act consequentialism', so that non-maximizing theories, such as those concerned with satisficing, can be accommodated under the 'act consequentialism' label. But I shall follow other writers (Sen, p. 464; and Scheffler, p. 1) in using the term in this way.

Act consequentialism is a moral theory. Its analogue among theories of rationality is *egoism*:

> An act is rational if and only if its consequences are at least as good for the agent as those of any other act that the agent could have performed instead.

Again, terminological questions may be raised; in particular, it might be asked why this view is not called *act* egoism. But whereas the term 'consequentialism' is understood in ordinary use to include rule-based and other indirectly maximizing theories, the term 'egoism' is not generally construed so broadly.

Act consequentialism and egoism obviously pertain to different domains: morality and rationality, respectively. But they also have an obvious structural similarity: they both require

[6] Exactly what makes a theory consequentialist or not is a surprisingly vexing question, and one that I shall not try to answer, or even to explore, here. For some discussions of it, see Stocker (1969, p. 276), Kupperman (1980, pp. 327–330; and 1981, pp. 305–306), Parfit (1984, pp. 26–27), Slote (1984, pp. 140–144; and 1985, pp. 35–39), Kagan (1989, p. 7), Pettit (pp. 230–231), Griffin (1992, pp. 118–120 and p. 125; and 1995, p. 154), Blackburn (1994, p. 77), Scarre (pp. 10–14), Gaut (p. 176), and Shaw (1999, p. 12 and p. 75).

[7] Admittedly, some writers use the term 'utilitarianism' to refer to a theory that is already as general as act consequentialism. See, for example, Lyons (1965, p. vii), Hodgson (p. 1), and Regan (p. 1).

For some accounts of the features that distinguish utilitarian theories from other consequentialist moral theories, see Rawls (1999b, pp. 22–23), Kagan (1992, p. 233), Scarre (pp. 4–26) and Shaw (1999, pp. 11–12). For an account of the different variants within the class of utilitarian theories, see Casteñeda: he purports to find 49 of them (p. 257).

agents to choose acts that maximally advance certain aims. On the basis of this similarity, we shall refer to both of these theories as *straightforwardly maximizing* normative theories.[8]

3.2 As alternatives to straightforwardly maximizing theories, some writers have proposed what we may call *indirectly maximizing* theories. Some of these are versions of rule consequentialism, a standard version of which says that

> An act is right if and only if it would be allowed by rules whose general acceptance would have better consequences than the general acceptance of any other rules.[9]

So whereas the act consequentialist identifies the right act (for a specific situation) by surveying the consequences of all of the possible acts, the rule consequentialist identifies the right act by surveying the consequences of the general acceptance of all of the possible rules, selecting those rules whose acceptance has the best consequences—which we may call the *optimal* rules—and finding an act, or acts, that comply with those rules. Analogously, a standard form of rule egoism would invoke an *agent's* optimal rules in holding that

> An act is rational if and only if it would be allowed by rules whose acceptance by the agent would have better consequences for him than his acceptance of any other rules.[10]

Note, then, that rule egoism is not a version of what is here regarded as egoism; it is an alternative to it. Moreover, revisionist theorists of rationality, such as Gauthier and McClennen, have advanced other, non-rule-based, indirectly maximizing theories of rationality, such as constrained maximization and resolute choice, that we shall consider in more detail later.

[8] It could be argued that the apt term here is not 'maximizing', but 'optimizing'. See, for example, Rescher (1993, ch. 2 [pp. 26–44]). But since all of what follows could easily be restated in terms of optimization, I shall employ the familiar (from the literature) terminology of maximization.

[9] A similar principle is formulated by Hodgson (p. 166).

[10] Similar principles are offered by Stanley Moore (p. 45 and p. 48), Brandt (1972, p. 691), and Kavka (1986, p. 358–59). But these writers offer rule-egoistic *moral* theories, on which an act is *right* if and only if it would be allowed by rules whose acceptance by the agent would have better consequences for him than his acceptance of any other rules. Hospers develops yet another rule-egoistic moral theory: "one should observe *those rules whose adoption* [throughout one's society, not just by oneself] *would be to one's interest*" (p. 393). Another variant is suggested by Kalin (p. 339).

Also note that several recent commentators have converged on the view that Hobbes's moral theory is a rule-egoistic one. See, for example, Stanley Moore (pp. 45–46), Kavka (1986, pp. 360–363), Gauthier (1987, p. 268), and Mulholland (p. 548). Hollis, however, appears to revive the interpretation of Hobbes as a divine-law theorist in his claim that although Hobbes "seemed to offer either to dispense with moral obligation as the basis of trust or to ground obligation squarely in rational self-interest," Hobbes "then tries to deliver the genuine moral article by invoking God as a joker" (p. 35). Despite the possibility of plausible rule-egoistic theories of *morality*, the only rule-egoistic theories we shall consider are ones of *rationality*.

3.3 The theories with which we shall be concerned, then, may be categorized in the following way:

		morality	rationality
straightforwardly maximizing		act consequentialism	egoism
indirectly maximizing	rule-based	rule consequentialism	rule egoism
	non-rule-based		constrained maximization, resolute choice

The cell for non-rule-based indirectly maximizing theories of morality is empty because the only such theories with which we shall be concerned are the moral-theoretic analogues of constrained maximization and resolute choice.

4 Decision procedures

4.1 One notion that will be employed throughout this dissertation is that of an agent's accepting a normative theory as a *decision procedure*. To understand this notion, let us take egoism as an example. For an agent to agent to accept egoism as his decision procedure is for him to take the production of the best possible consequences for himself, or the maximal advancement of his interests, as determinative in his deliberations. That is, it is for him to seek in each of his acts to make his life go, from that point on, as well as possible.

This does not mean that he must start from scratch in his estimation and comparison of consequences. On the contrary, he may use rules of thumb, or other guidelines summarizing the reasoning about causes and effects made available to agents by their own and others' accumulated experience. An egoist may, for example, follow his doctor's orders when sick, may follow recipes when cooking, and may follow other instructions and guidelines as he sees fit. But, *qua* egoist, he must (1) use such rules only in order to ascertain or estimate what conduct will maximally advance his aims, instead of treating them as "the last word" for the situation in which he finds himself, and (2) set them aside when he believes that doing so would advance his interests more than following them would. He may not, that is, regard himself as rationally bound to follow the rules when he judges them inapplicable or simply mistaken. This is not to say, of course, that an egoist must have no respect for such rules; on the contrary, *qua* egoist he may be quite deferential to such rules, insofar as he regards himself as not qualified to judge the rules inapplicable or mistaken. Egoists may display epistemic timidity, but not deliberative timidity. The same, of course, goes for act consequentialists.

4.2 As I said, we will be concerned throughout this dissertation with agents' accepting normative theories as decision procedures. In particular, we will be concerned with the *consequences* that result from such acceptance. Some theorists, when writing about the consequences of various theories' being put into practice in this way, suggest that such an inquiry—into the consequences that result from such acceptance—is likely to be fruitless. Regan, for example, writes that

> So far as I can see, it is not possible to say very much at all about the consequences of agents' trying to satisfy, or accepting, various theories. . . . [T]he consequences of agents' satisfying theories seem to me more significant for the choice between theories than the consequences of agents' accepting theories, or whatever. (p. 2)

Regan, then, focuses on the consequences of agents' *complying* with certain theories, regardless of whether they actually accept those theories as decision procedures.

Now Regan's approach does have the important merit of abstracting from problems of agents' lack of information, imperfect calculation abilities, and insufficient moral motivation: problems of implementation that tend to arise if one focuses on agents' accepting certain theories as decision procedures. For implementation problems are only one of the several important phenomena that arise when agents accept certain theories as decision procedures, and it is crucial not to get hung up on them. But implementation problems are still important; moreover, there are some aspects of agents' accepting certain theories as decision procedures that it is vitally important not to overlook, and these would be obscured if not lost altogether if we focused compliance rather than acceptance. So I hope to gain the benefits of Regan's approach without paying the price Regan pays by focusing on acceptance, not compliance, but—in so doing—not attending unduly to those consequences of acceptance that are not also consequences of compliance.[11]

[11] The propriety of focusing on acceptance, not compliance, is argued with more thoroughness by Hooker (2000, pp. 75–80).

5 Standards

5.1 I said in the opening paragraphs of this chapter that we ought to focus on standards, not theories; and yet most of this chapter has been about theories. It is time now to turn to standards and to elaborate on our strategy of seeking insight into the merits and demerits of theories by focusing on the standards used to evaluate them. For this strategy is not a typical one. Indeed Regan writes that

> So far as I am aware, no one has tried to identify explicitly the general properties we would like a utilitarian theory to have. As a result, no one has attempted to produce a systematic analysis telling us which theories have which desirable properties, and considering whether various kinds of theories can have various desirable properties, singly or in combination. (p. viii)

Although we shall not attempt to cover all of the territory that Regan mentions, we may hope to break some new ground there.

To be sure, some moral theorists have broached the subject of such standards. Shaw, Hooker, and Hare,[12] for example, provide some lists of standards that theories ought to meet. But each author presents his standards as a *settled* list of requirements (or, at least, desiderata)—not as objects of evaluation and comparative assessment in their own right. Some scrutiny of standards themselves is offered by Jane Singleton, who aims to show that "more attention needs to be paid to the question of tests of adequacy for moral theories, and that these should be made clear before the theories are formulated in detail" (p. 31; cf. p. 43). Her sentiment is laudable, but her scope is so wide—encompassing not only moral theories but also meta-ethical theories—that the strongest conclusion her inquiry yields is that "there is a fundamental difference between the

[12] Shaw writes that some particular conception of morality could not aspire to be "our morality" unless it is "theoretically plausible," "intellectually attractive," "psychologically tenable," and "socially feasible" (1980, p. 133). He adds that its principles must be "teachable," it must be "reasonable for an individual or society to endorse it," the "psychological strains of adhering to it" must not be too severe, and people must be "able to comply with it successfully" (1980, p. 133). Finally, he says, a moral code must not be "inconsistent, too difficult for people to master, psychologically untenable, or whatever" (1980, p. 134). There is some overlap, then, between the standards at which Shaw gestures and the ones we shall examine.

Hooker writes that a moral theory should start from attractive general beliefs about morality, be internally consistent, have implications that match our considered judgments, provide a fundamental principle that unites and justifies our considered judgments, and help us deal with moral questions about which we are not confident, or do not agree (2000, p. 4; see also 1996, p. 531). Hooker's standards basically add up to the approach to moral-theory selection known as reflective equilibrium, which falls outside the purview of this dissertation.

Finally, Hare writes that he hopes to identify the best ethical theory by "mak[ing] a list of the requirements that an adequate ethical theory has to satisfy. Then we can look at each [candidate] theory in turn and see which of the requirements it satisfies, and which it fails to satisfy" (1997, p. 46). Unfortunately, what Hare means by an ethical theory is not a normative theory, but a meta-ethical one, so the requirements he specifies are not ones that we can expect to be relevant to our inquiry.

tests of adequacy that are required for first order and second order moral theories" (p. 46). An inquiry such as ours—which focuses on a handful of standards used to evaluate just *normative* (or first-order) theories—may be expected to yield stronger conclusions, or at least deeper understandings of the standards to be examined.

5.2 Another issue that arose in the opening paragraphs of this chapter is the possibility that an inquiry such as this one might seem suspiciously abstract, or pointlessly removed from the substantive questions of morality and rationality. By now it should be clear that this kind of inquiry is, in fact, foundational, since no theory offering answers to those substantive questions can be accepted until we know how to evaluate such theories. And once we know how to evaluate theories, then many of the debates about theories themselves may turn out to be superfluous. As Singer writes,

> The criterion by which we decide to reject, say, utilitarianism in favour of a contractual theory of justice (or vice versa) is, if anything, even more fundamental than the choice of theory itself, since our choice of moral theory may well be determined by the criterion we use. (1974, p. 490)

Clearly, any results obtained in a study of standards would have ramifications for specific theories. So the question of standards is not an optional tangent from, but rather an integral component in, the development of acceptable normative theories.

Ultimately, then, this dissertation is meant as a contribution to the ongoing debates within moral philosophy and rational-choice theory about the relative merits of various normative theories. The precise nature of this contribution will not be to resolve any of these debates to the satisfaction of all of the parties to it, but will be (in part) to show the necessity of shifting the debate to the level of the standards used to evaluate theories, instead of remaining content to point out the apparent pros and cons of the theories themselves.

6 A look ahead

The organization of this dissertation is fairly simple. The long chapter II is essentially expository, specifying what it means (for the purposes of this dissertation) for a normative theory to be self-defeating and setting out some of the many sources of the self-defeat of straightforwardly maximizing theories such as act consequentialism and egoism. These results motivate the first of the three standards to be examined: that a theory not be self-defeating in the specified sense. The merits of this standard are investigated—and found wanting—in chapter III. As we shall see there, a standard that naturally arises in the debate over self-defeat is that a theory not violate the publicity condition; chapter IV is devoted to explaining and debunking this standard. In chapter V, I explore and defend the claim, often made against indirectly maximizing theories such as rule consequentialism and revisionist theories of rationality such as those of Gauthier and McClennen, that a theory must not be incoherent in a way that those prove to be.

It will not have escaped the notice of the discerning reader that the cumulative effect of these chapters, as just described, will be to offer an unqualified defense of straightforwardly maximizing theories and a wholesale indictment of their indirectly maximizing rivals. I shall address this point in the concluding chapter VI. There I shall also offer some remarks on the possibility—or, as I shall suggest is more likely, the *impossibility*—of avoiding taking sides by formulating a theory that meets all of the standards in question.

II

Sources of Self-Defeat

> some of the Tory leaders . . . quoted against me certain passages of my writings and called me to account for others, especially for one in my *Considerations on Representative Government* which said that the Conservative party was by the law of its composition the stupidest party. They gained nothing by drawing attention to this passage, which up to that time had not excited any notice, but the *sobriquet* of 'the stupid party' stuck to them for a considerable time afterwards.
>
> —John Stuart Mill (1873), p. 212

1 Introduction

In this passage of his *Autobiography*, Mill offers an amusing example of a familiar phenomenon—a phenomenon familiar enough, in fact, to have earned at least a couple of clichés: a plan backfires; a person shoots himself in the foot. The Tories' efforts were, one could say, self-defeating.

We know what it means to say that an act or a course of action is self-defeating. But what does it mean to say that a normative theory is self-defeating? And what would be the implications of such a claim for our evaluation of that theory? We normally think that it's an occasion for criticism if an act or a plan or a policy is self-defeating. Is it also an occasion for criticism if a normative theory is?

The sense in which self-defeat may apply to a normative theory closely parallels the sense of self-defeat found in ordinary conversation. Jon Elster, in his works, offers many examples of self-defeating activities; here is a small sample of his findings:

> Concern with outcomes can be self-defeating. . . . Insomnia, impotence and stuttering get worse if one tries to do something about them. . . . . Spontaneity will elude us if we try to behave spontaneously. . . . We may wish to be esteemed and admired by others, but actions that we undertake for the sole purpose of achieving this end will undermine themselves. (1989, p. 24)

Typically, we speak of individual acts as self-defeating. But entire policies can be self-defeating, too. In *Silent Spring*, Rachel Carson describes a policy of spraying a certain pesticide in certain areas. The pesticide not only failed to kill the pests for which it was intended; it also killed the pests' natural predators (pp. 245–261). The policy of spraying this pesticide in these areas was plainly a self-defeating one.

In exactly what sense, then, can a normative theory be self-defeating? We shall regard a normative theory as self-defeating if the outcomes at which it aims would be worse achieved if agents adopted and accepted that theory as their decision procedure than if they adopted and accepted some other theory as their decision procedure. That is, we shall construe self-defeat in terms of *pragmatic ineffectiveness*.[1]

Many theories may be self-defeating in this sense. But straightforwardly maximizing theories are the ones of which this is said most often (probably most often *by far*), and they shall be our focus in this chapter. That is, we shall survey the many ways in which such theories may be self-defeating: the many sources of self-defeat that may lie within them. The sources of self-defeat to be considered are of three kinds. The first of these, which has to do with the implementation of straightforwardly maximizing theories, is the subject of section 2. In section 3, we consider the ways in which normal human desires, along with the demands of normative theories, can impede the attainment of the best outcomes. Finally, in sections 4 through 7, we shall consider self-defeat as it arises from dynamic inconsistency. The following table shows, in somewhat greater detail, how various sources of self-defeat (sources of pragmatic ineffectiveness) are allocated among the sections:

			section
implementation			2
human desires and theoretical demands			3
dynamic inconsistency	non-strategically induced		4
	strategically induced	individual	5 and 6
		collective	7

[1] Note, though, that the pragmatic ineffectiveness in question is ineffectiveness in the achievement of only those outcomes aimed at by the theory. A theory may be pragmatically ineffective in the achievement of some outcome, without thereby being self-defeating, if that outcome is not one at which the theory aims—either because it aims only at other outcomes or because it aims at no outcomes at all.

Before proceeding I should mention that within the discussion of each kind of self-defeat, I shall be starting with egoism and then moving on to act consequentialism, since the one-person problem of instrumental rationality is simpler in certain respects than the multi-person problem of morality.

2 Implementation: easier said than done

2.1 In this section, we shall consider one class of factors that may cause straightforwardly maximizing theories such as egoism and act consequentialism to be self-defeating: the class of factors that impede the successful implementation of these theories. These sources of self-defeat are generally regarded as less significant than sources of the other two types to be discussed (in section 3 and in sections 4 through 7) but they are worth noting at this stage of our inquiry.

2.2 For the egoist, the challenges of implementation begin with the fact that it is often exceedingly difficult for an agent to notice all of the acts open to him. Moreover, even when an agent can identify all of his options, it is likely to be exceedingly difficult for him to identify all of the ways in which his options' consequences would affect him.[2] Third, even when an agent can identify all of the effects on him of all of his options' consequences, it may be difficult for him to assess their relative merits, in terms of how these effects on him would contribute or detract from his interests. These observations, of course, are not new; readers familiar with the literature on consequentialism will recognize these points as well-established ones. Indeed more than a century ago Sidgwick could already say that "there is scarcely any point on which moralisers have dwelt with more emphasis than this, that man's forecast of pleasure is continually erroneous" (p. 142), and it seems fair to say that they could have dwelt with equal justice on an equally unflattering assessment of man's forecast of other aspects of his existence.

A fourth problem of implementation for the egoist is that his judgment may be clouded by the temptation to perform an act whose benefit to him now, or in the immediate future, is greater than that of the act whose *overall* benefit to him is greatest. The temptation I am referring

[2] Consider, for example, the difficulties involved in following the familiar advice, "Buy low, sell high." Also, reflect on the fact that everyone has found that seemingly insignificant events in one's life can have a variety of unforeseen consequences, some of them quite substantial. It follows that in order for an egoist to enjoy even the most modest level of pragmatic effectiveness, he must be gifted with an acute sensitivity to this phenomenon and an ability to predict its unfolding with remarkable accuracy.

to is not simply a matter of the agent's engaging in time discounting—in other words, not simply a matter of the agent's failing, or refusing, to regard all parts of his life as equally important. For although Sidgwick maintains that

> this equal and impartial concern for all parts of one's conscious life is perhaps the most prominent element in the common notion of the *rational*—as opposed to the merely *impulsive*—pursuit of pleasure (p. 124, n. 1)

it seems unnecessarily restrictive to agree with this remark's implication that time discounting is, in itself, disallowed by legitimate canons of instrumental rationality (regardless of how important it is to, as Sidgwick called it, the "common" notion of the rational). Indeed Elster writes that policymakers are "open to criticism" if they refuse to incorporate their constituents' time discounting into policies on the grounds that the task of the policymaker is to take a more farsighted view (2000, p. 165). So the temptation I am referring to is not time discounting itself. Rather, the temptation I am referring to is the phenomenon—recognizable at least conceptually, if not often in practice—of an agent's giving his present self, or his immediately succeeding selves, more priority than even his time-discounted conception of the good life for himself would allow.

Fifth, even if the agent is able to choose the optimal act, he will often be able to do so only after devoting considerable time and mental resources to identifying it. Indeed we may make the stronger claim that he will often be able to do so only after devoting so much time and mental resources to identifying it that he thereby loses more in calculation costs than he gains from choosing the optimal act instead of the act he would have chosen if he had employed some other, less calculation-intensive, decision procedure.

These five problems associated with the implementation of egoism suggest that the adoption of egoism as a decision procedure by a normal human agent would have worse consequences than the employment by him of some other normative theory as decision procedures. Of course, we cannot regard this suggestion as conclusively established until we have identified a rival that would be more pragmatically effective than egoism; it is not enough just to find that the results of an agent's being an egoist are worse than one might have hoped. But the foregoing considerations are a start. They help us to see, at least, why Sidgwick writes that "A dubious guidance to an ignoble end appears to be all that the calculus of Egoistic Hedonism has to offer" (p. 200).

2.3 The foregoing remarks may easily be adapted to apply to act consequentialism. Like egoism, act consequentialism would be forbiddingly difficult to put into practice. The first three reasons offered for this claim in regard to egoism—the difficulty of noticing all of one's options, the difficulty of ascertaining their consequences, the difficulty of ascertaining those consequences' merits—may easily be transferred to the present case. But the third problem is exacerbated by the agent's need to assess not only his options' morally relevant effects on *him*, but also his options' morally relevant effects on everyone in the scope of the act-consequentialist theory he is using (be it all humans in his society, all humans everywhere, all sentient creatures everywhere, or whomever). And the fourth problem of implementation for egoism (excessive priority for one's present self or immediately succeeding selves) finds its analogue for act consequentialism in the problem of an agent's being tempted to perform an act whose benefit to individuals (including himself) now, or in the immediate future, is greater than that of the act whose benefit to individuals *generally* is greatest. Moreover, a problem for act consequentialism related to this fourth one, but without (as far as I can see) an analogue in the case of egoism, is the familiar problem of an agent's being tempted to perform an act that benefits him more than the optimal act does—especially if the identity of the optimal act is cloudy for any of the first three reasons (Parfit 1984, p. 28). As Brandt writes,

> the problem to be solved [by the agent] can be complex enough to invite self-serving rationalization. Suppose one is considering whether to make an income tax report that dodges payment of $1,000. The benefit is partial payment of a vacation in Greece for me. What is the cost in general utility? Who will have which benefit, and who would not have it if I save money for a vacation in Greece? (1996, p. 143)[3]

Finally, the fifth problem discussed in the case of egoism (excessive costs of calculation) carries over without amendment as the sixth problem in the case of act consequentialism. These considerations suggest that act consequentialism would be at least as self-defeating as egoism would be. Indeed Sidgwick concludes that "from the universal point of view no less than from that of the individual, it seems true that Happiness is likely to be better attained if the extent to which we set ourselves consciously to aim at it be carefully restricted" (p. 405).

[3] For another, more vivid (and, thus, more distressing) example, see the ninth chapter of George Eliot's *Romola*, in which Tito satisfies himself that he needn't use the funds he has obtained by selling his father's gems in order to try to rescue him from the slavery he believes him to be in, because he's not *sure* he's alive, he's not *sure* he can find him, and he's not *sure* he can make the necessary voyage safely.

2.4 Although I shall defer until the next chapter a general assessment of the significance of self-defeat, I want to pause to address the claim that implementation problems, in particular, are not a valid basis on which to criticize egoism and act consequentialism. How serious is self-defeat, when due to such factors? Mill suggests that it is not serious at all when he points out that “There is no difficulty in proving any ethical standard whatever to work ill, if we suppose universal idiocy to be conjoined with it” (1861, p. 224 [ch. II, par. 24]). And we may obviously extend Mill’s remark to cover theories of rationality as well as moral theories. But as forceful as this reply may seem, it does not succeed in trivializing every instance of self-defeat due to implementation problems, since not every such instance of self-defeat arises from conjoining *universal idiocy* to the theory or theories under consideration. Indeed what we have assumed in our discussion is not “universal idiocy,” but simply the capacities for means-end reasoning and resisting temptation that normal human beings have. In the cases of egoism and act consequentialism these capacities may bear, insofar as their effects are concerned, an unhappy resemblance to utter incompetence; but there may be other theories of rationality and morality that are easier (both intellectually and motivationally) for agents to apply, such that the capacities for means-end reasoning and resisting temptation that normal humans have would be fully adequate for the successful, and pragmatically effective, implementation of the theory.

3 Human desires and theoretical demands

3.1 We have, then, considered several sources of self-defeat that may be found in the category of implementation problems. In this section, we shall consider cases in which implementation problems do not arise. In such cases, an agent successfully chooses and performs the act that is optimal for him (in the case of egoism) or optimal overall (in the case of act consequentialism), and does so without substantial calculation costs. But even when such competence is stipulated, egoism and act consequentialism may nevertheless exhibit self-defeat due to certain psychological features that agents normally possess.

3.2 One reason why even effortlessly implemented egoism might be self-defeating stems from the fact that for most people, the happiness of others is an important ingredient in one’s own happiness. And it is typically a sufficiently important ingredient that an agent does better if he cares about others for their own sakes than if he cares about them merely as creatures to be made happy in order to enhance his own happiness. As Hodgson writes,

> It may well be that certain of the most happy experiences which persons can enjoy are open only to those who care about other persons for their own sakes. For example, a person who cared only about his own happiness could not be happy simply because of the happiness of other persons; and it seems likely that he could not fully enjoy close personal relationships. (p. 61)

On the basis of reasoning of this sort, Hodgson concludes that "even if an egoist was successful in always choosing the course of action which in the circumstances would give him greatest happiness, he might be less happy than if he was not an egoist" (p. 61).[4]

Moreover, even if a person's happiness does not depend on others', his own happiness will probably be better promoted if he pursues activities and projects as if they were worth doing for their own sakes, not merely as sources of happiness for himself.[5] This thought led Mill to write that although he "never, indeed, wavered in the conviction that happiness is the test of all rules of conduct, and the end of life," he came to realize that

> this end was to be attained by not making it the direct end. Those only are happy (I thought) who have their minds fixed on some object other than their own happiness. . . . Aiming thus at something else, they find happiness by the way. (1873, p. 117)

We noted in section 1 of this chapter that Elster has observed the phenomenon of self-defeat in many areas of life; not surprisingly, one of his topics is the pursuit of happiness itself. He echoes Mill's sentiments in his remark that

> The individual who sets out to obtain pleasure or to make himself into a cultured person will usually be thwarted, unless at some point the means become ends in themselves. (1984, p. 40)[6]

So, certain facts of human psychology—such as the dependence of persons' happiness on that of others and the superior satisfactions to be derived from activities and projects pursued for their own sakes—suggest that the deck is stacked against the egoist from the start.

[4] For further discussion of this point, see Stocker (1976, p. 456) and Railton. The latter refers to it as the "paradox of hedonism" (pp. 140–141).

[5] For an example of such a case, see Parfit's discussion of Kate, the writer (1984, p. 6).

[6] See also his remark that "the conscious and deliberate attempt to maximize [one's own] utility tends to be self-defeating. It is a truism, and an important one, that happiness tends to elude those who actively strive for it" (1983, p. 9).

3.3 The facts of human psychology may undermine the pragmatic effectiveness of act consequentialism as well. One problem is that act consequentialism may be very demanding. As Brandt writes,

> the theory seems to make harsh and oppressive demands on the moral agent. For instance, at the present time I am relatively certain that, come evening, all my moral obligations (except the duty to give more to needy causes) will be discharged, and indeed that I will be free to go on writing, or to read a book or watch TV, without being morally derelict. But if I really have an obligation to do what will maximize benefit, this evening I might be obligated to phone some acquaintance and find out if there is something I could do to improve the quality of his life (provided there was no reason to think he would be irritated by the interruption). Doubtless it would be better if I felt called upon to do more of this, particularly for needy people, but the idea of always being morally obligated, with no area of freedom to do just what I want or to enjoy myself (except where this would maximize expectable utility), is not a recommendation to me. (1996, p. 144)

Indeed if we were act consequentialists then we might feel so beset by feelings of duty and obligation that our lives would be much less pleasant than they are. The mental life of a person constantly consulting the act-consequentialist moral standard, in order to figure out what to do next, would no doubt be an unpleasant one.

Moreover, even if agents were successful in meeting the demands of act consequentialism, a new problem would be created. To see this, begin with Parfit's observation that

> Most of our happiness comes from having, and acting upon, certain strong desires. These include the desires that are involved in loving certain other people, the desire to work well, and many of the strong desires on which we act when we are not working. (1984, p. 27)

Now anyone who is an act consequentialist would have to act against these desires from time to time, and if (as we are assuming) he were successful, then these desires would almost certainly be dampened over time, denying the agent an important source of happiness. Parfit concludes that in such a world we might always do what would have the best consequences, but the outcome would still be worse than if we had certain other desires and dispositions (1984, pp. 27–28).[7]

[7] See also Stocker's claim that "To the extent that you live the theory directly, to that extent you will fail to achieve its goods" (1976, p. 461). For criticism of Parfit's argument, see Mendola and Moser.

4 Dynamic inconsistency, part 1: non-strategically induced

4.1 So much for two of the three broad kinds of sources of self-defeat for straightforwardly maximizing theories such as egoism and act consequentialism. In order to focus on the third—dynamic inconsistency—we abstract from cases such as those discussed in the last two sections by assuming that agents are both (1) able to implement egoism or act consequentialism without cost and (2) constituted in such a way that happiness for them neither depends on that of others nor is enhanced by the pursuit of activities and projects for their own sakes. Such individuals would be strange indeed, and it may seem inevitable that considering such agents would plunge us into fruitless abstractions. But in fact we shall find in this context the most remarkable instances of self-defeat.

What does it mean to say that egoism or act consequentialism may give rise to dynamic inconsistency? It means that an agent who adopts such a theory as his decision procedure may thereby be led at one point to plan on a course of action (or to wish to plan on a course of action) that his decision procedure will later counsel him, or would later counsel him, to abandon. The cases of interest, then, are ones in which either (1) when the agent is part of the way down a road his decision procedure counseled him to take, his decision procedure then counsels him to diverge from it or (2) the road that he would like to choose is not one that he can regard as a genuine option because of his anticipation that his decision procedure would counsel him to diverge from it if he were somehow to choose it.

Elster notes that cases of dynamic inconsistency may be divided into two main classes: that caused by time discounting and that caused by strategic interaction. He writes that

> Apart from a certain formal similarity, the two have little in common. [Time] discounting does not require interaction: it might apply to Robinson Crusoe on his island before the arrival of Friday. Conversely, strategically induced inconsistency does not require discounting. . . . [T]he two phenomena can interact, but either can exist without the other. (2000, p. 24)

It will be convenient to organize our own findings along the lines of Elster's distinction, since the cases of non-strategically-induced inconsistency can naturally be separated from the cases of strategically induced inconsistency. In this section, we will focus on cases of non-strategically induced inconsistency, deferring cases of strategically induced inconsistency until sections 5 through 7.

4.2 In order to see how egoism can lead to dynamic inconsistency of the sort that gives rise to self-defeat (still understood, of course, as pragmatic ineffectiveness), consider some results from a seminal paper by R. N. Strotz.

Strotz asks us to imagine an agent with a fixed stock of some resource that he consume in the rest of his life. Knowing (let us suppose) how much of this resource he has and how much longer he will live, the agent selects a consumption plan. As noted earlier (in subsection 2.2), this consumption plan needn't treat the agent's successive future selves equally; time-discounting is permitted. But his plan does take his future selves into account. Now it is natural to think that whatever plan the agent settles on will be such that if he later reconsiders his situation and considers his choice anew—taking fresh stock of the amount of the resource he has left and the time he has left—then he would settle on a plan that is consistent with the plan he originally selected. Strotz, however, establishes the surprising result that this may not be the case.

We can see the logic behind Strotz's result by way of a simple numerical example. Suppose all of the following:

1. I am an egoist. Whenever I have a choice to make in the realm of individual, instrumental rationality, I optimize.
2. I know, today, that I will get to choose, tomorrow, between then enjoying a benefit that will be worth 10 units of utility to me (option A) and declining that option in order to enjoy, the next day, a benefit that will be worth 20 units of utility to me (option B)—as shown in Figure II.1. I want to decide, today, which option I am going to choose, so that I can plan today's activities accordingly.
3. My time discounting is hyperbolic: that is, the present value to me of a benefit with value *X*, to be enjoyed *t* periods of time from now, is $\frac{X}{1+kt}$, where $k > 0$ (Elster 2000, p. 25). The greater *k* is, the sharper is my discounting. Obviously what *k* is for anyone will depend on how long a time period is. Let us say that a time period is one day long and that when this is so, then $k = 1.5$ for me.

Given these assumptions about me and the situation in which I find myself, the present value of option A, which is one period away, is $\frac{10}{1+1.5}$, or $\frac{10}{(5/2)}$, or 10 times $\frac{2}{5}$, or 4. (In Figure II.1, this is shown by the height, when time = 1, of the curve leading to point A.) And the present value of option B, which is two periods away, is $\frac{20}{1+3}$, or $\frac{20}{4}$, or 5. (This is shown by the height,

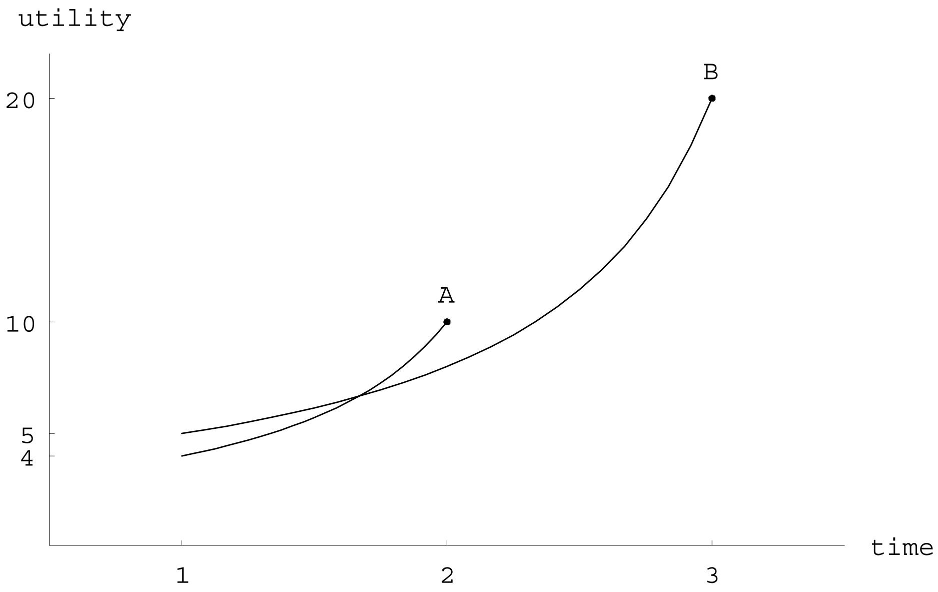

Figure II.1

when time = 1, of the curve leading to point B.) So today I deem option B to be preferable to option A. But what about tomorrow, which is the time for choosing (taking option A, in which case I'll be denied option B the next day, or declining option A, in which case I'll enjoy option B the next day)? Tomorrow, option A will not be discounted, so its present value will simply be 10; and option B will have a present value of $\frac{20}{1+1.5}$, or $\frac{20}{(5/2)}$, or 20 times $\frac{2}{5}$, or 8. (So, in the figure, when time = 2, point A is higher than the line leading to point B.) My preferences will have reversed: option A will have a higher present value, and since it will be time to choose, I will take it instead of (as I would plan today to do tomorrow) waiting for option B. I prefer option B today, and option A when it is time to choose. I am dynamically inconsistent.[8]

[8] It may be objected that few agents would have discounting as steep as that implied by $k = 1.5$. And of course this value for k, along with the utilities associated with benefits A and B, was chosen with a view to constructing this example. But the example can be generalized in two significant ways. First, no matter how small k is (as long as it is positive), and no matter what the value of option B is (call it b), there exists a range of numbers such that if the value of benefit A is in that range, then on day 1 I will prefer B-on-day-3 to A-on-day-2 and yet on day 2 I will prefer A-on-day-2 to B-on-day-3. (That range is from f to $f + \frac{bk^2}{2k^2+3k+1}$ where $f = \frac{2kb+b}{2k^2+3k+1}$.) Second, no matter how small k is (as long as it is positive), and no matter what the values of benefits A and B are (call them a and b), there exists a range of numbers such that if t is in that range, then on day 1

Of course, hyperbolic discounting is not the only sort of time discounting that can give rise to dynamic inconsistency; moreover, time discounting is not the only source of non-strategically induced dynamic inconsistency. For example, an agent may also be led into dynamic inconsistency by violating a principle of choice known as the independence principle, as McClennen shows (1990, pp. 6–11).[9] But the foregoing example is enough to show how even an agent who infallibly optimizes at every choice point can be led to an outcome that he regards as non-optimal.

4.3 The dynamic inconsistency of egoism just exposed has an obvious analogue in the case of act consequentialism. To modify the foregoing example in order to make it relevant to act consequentialism, it is sufficient to suppose that I am an act consequentialist, instead of an egoist, and to suppose that options A and B have *overall* values of 10 and 20, not just those values *for me*. Then, when trying to choose the best outcome (for everyone, not just me), I suffer the same reversal of preferences and the same inability to forgo option A in order to obtain option B.

4.4 We have seen that an agent who chooses conduct with a view to advancing his (egoistic or act-consequentialist) aims does not further those aims as much as his circumstances permit. But this fact alone does not militate against straightforwardly maximizing theories. For an agent who furthers his aims as much as his circumstances permit, but who does so without choosing conduct for *reasons*—who does so, for example, through instinct, under hypnosis, or simply with the benevolence (real or metaphorical) of nature—does not exhibit the pragmatic effectiveness of a *competing* normative theory, because he does not exhibit *any* normative theory. Choice of the sort that interests us is choice made on the basis of *reasons*, and choosing conduct for straightforwardly maximizing reasons can be pragmatically ineffective only insofar as an agent who so chooses does not further his aims as much as he would if he were to choose conduct for *other* reasons.

Gauthier and McClennen, it is to be admitted without quarrel, supply other such reasons. The reasons proposed by Gauthier are those recognized by his constrained maximizer, who chooses conduct not solely on the basis of its optimality, but also on the basis of such reasons as

I will prefer B-on-day-*t* to A-on-day-2 and yet on day 2 I will prefer A-on-day-2 to B-on-day-*t*. (That range is from g to $g + \frac{b-a}{a}$ where $g = 2 + \frac{b-a}{ka}$.)

[9] See also McClennen (1990, pp. 151–152 and p. 167).

its compatibility with the plan or plans the forming and execution of which, considered as a single course of conduct, can be expected to further his ends more than they would otherwise be furthered (1975a, p. 227–229). This portrayal of constrained maximization, though capturing none of the detail of Gauthier's conception of it,[10] is sufficient for our purposes because our concern is with its appeal to reasons other than, or at least in addition to, the optimality of conduct. McClennen, too, argues for the validity of reasons other than the optimality of conduct; the ones proposed by him are those recognized by his resolute chooser. Like the constrained maximizer, the resolute chooser chooses conduct not solely on the basis of its optimality, but also on the basis of such reasons as its compatibility with a previously adopted plan: "the agent can be interpreted as resolving to act in accordance with a particular plan and then subsequently intentionally choosing to act on that resolve, that is, subsequently choosing with a view to implementing the plan originally adopted," even though "the plan . . . calls for . . . a choice that the agent would not otherwise be disposed to make" (1990, p. 157).[11] Although we will examine these theories of rationality in much more depth as this dissertation progresses, what is of immediate interest to us is that both the constrained maximizer and the resolute chooser can plan to choose on day 1, and can successfully choose on day 2, to decline option A on day 2 and enjoy the benefits of option B on day 3. Moreover, we can imagine moral-theoretic analogues of these indirectly maximizing theories of rationality that would be pragmatically superior to act consequentialism. Choosing conduct for reasons other than those of optimality, then, enables an agent to avoid certain losses and to secure the benefits that we would expect from truly optimizing choice.

4.5 But is this pragmatic ineffectiveness—which seems so alien to the idea of choice made with a view to achieving the best outcomes—genuinely indigenous to the world of straightforwardly maximizing theories, or has it been planted there by its opponents? In this and

[10] For some specifications and modifications of constrained maximization, see Gauthier (1975a, pp. 227–230; 1984a, pp. 264–267; 1990a, pp. 4–5; 1994, pp. 702–707; and 1998, pp. 48–53) and McClennen (1988, pp. 96–97).

[11] For other statements of the conditions warranting resolute choice, see McClennen (1988, p. 112; 1990, p. 13 and pp. 209–213; and 1997, pp. 231–33 and p. 241) and Gauthier (1997b, p. 8).

In understanding resolute choice it is important to distinguish McClennen's sense of 'resolute' from others that may seem intuitive, two of which have been specified by Korsgaard. In one paper, she writes that we might call "*resolute*" someone who "respond[s] much more readily and definitely to the consideration that something is an effective means to their end" (1986, p. 325). In another paper, she explores the possibility that we might "call somebody 'resolute' only when he pursues ends of which we approve" (1997, p. 233). Obviously an agent could be resolute in either or these senses, or in any of many others, without being a McClennenian resolute chooser.

the next two subsections, I want to consider, and to reject, three replies that defenders of egoism and act consequentialism might make to the foregoing allegations of pragmatic ineffectiveness.

First they might point out that the sort of straightforwardly maximizing agent we have been considering—who handles the situation under discussion so ineptly—does not represent egoism or act consequentialism in its best light; and that a sophisticated straightforwardly maximizing agent will fare better. Drawing on Strotz's claim that "there is nothing patently irrational about the individual who finds that he is in an intertemporal tussle with himself—except that rational behaviour requires he take the prospect of such a tussle into account" (p. 171), a defender of straightforwardly maximizing theories might claim that a sophisticated egoist would not fall prey to inconsistency. As Strotz develops his suggestion,

> An individual who because he does not discount all future pleasures at a constant rate of interest finds himself continuously repudiating his past plans may learn to distrust his future behavior, and may do something about it. Two kinds of action are possible. (1) He may try to precommit his future activities either irrevocably or by contriving a penalty for his future self if he should misbehave. This we call the *strategy of precommitment*. (2) He may resign himself to the fact of intertemporal conflict and decide that his "optimal" plan at any date is a will-o'-the-wisp which cannot be attained, and learn to select the present action which will be best in light of future disobedience. This we call the *strategy of consistent planning*. (p. 173)

Such an agent would indeed be an improvement on the naïve egoist who plans to choose option B and then is surprised by his choosing option A. But the two parts of this reply still are not enough to answer the charge of pragmatic ineffectiveness.

To take the second part first: the strategy of consistent planning amounts to resigning oneself to the fact that one's optimal plan is unattainable. This strategy is, in effect, a denial of the very plausible idea that a rational agent can plan, today, to do something tomorrow, over and above merely anticipating that he'll do, tomorrow, whatever he would have done if he had not given the matter any thought today. But even if this pessimistic conclusion were warranted, it would leave unanswered the charge of pragmatic ineffectiveness. For an agent who employs the strategy of consistent planning, however thoroughly, still ends up with the inferior outcome. His only advantage over the naïve agent is that he foresees his pragmatic ineffectiveness—hardly a triumph of instrumental reasoning.

Nor does the other strategy, that of precommitment, rival (in terms of pragmatic effectiveness) constrained maximization or resolute choice. One problem, of course, is that precommitment is not always possible: we cannot always set up external structures in which we "deposit our will" (Elster 1984, p. 43) so that the desired consequences of the required acts can be effected without subsequent action by us. But even when it is, it has costs: tying oneself to the mast (either literally, like Ulysses, or figuratively, in any of a number of ways) consumes resources, as does contriving a penalty for one's future self if one should deviate from the plan one makes today. And side bets, if they do not consume resources, at least make them temporality unavailable, insofar as one's share must be placed in escrow.[12] Moreover, as Gauthier points out, the strategy of precommitment "fails to face the real issue—that taking my reasons for acting directly from my aim [which is how the egoist deliberates] is in certain situations counter-productive and, indeed, self-defeating in relation to that aim" (1994, p. 696). So even if the straightforwardly maximizing agent could ensure plan-consistent choice tomorrow, he would still do worse than his indirectly maximizing rival.

4.6 Second, it might be claimed that the source of the agent's pragmatic ineffectiveness in a case such as the one discussed above is not the agent's egoistic decision procedure, but the agent's time discounting—and not in the mere fact that the agent engages in *some* time discounting, but that the agent engages in dynamically inconsistent discounting. It may be pointed out that there are forms of time discounting that would preclude preference reversals in cases such as these, no matter what the relative magnitudes and temporal spacing of the goods are. Consider, for example, exponential discounting. If my discounting is of this form, then the present value to me of a benefit with value X, to be enjoyed t periods of time from now, is Xr^t, where $0 < r < 1$ (Elster 2000, p. 25). Say that when a time period is one day long, then r, for me, is 0.9. Then the present value to me of option B is 20 times $(0.9)^2$, or 20 times 0.81, or 16.2; and the present value to me of option A is 10 times $(0.9)^1$, or 9. So I want today to choose option B tomorrow. I still want this tomorrow, when option A is available without delay, since then the present value of option A, which of course is 10, is still less than that of option B, which is 18. As Strotz points out, exponential discounting saves the agent from dynamic inconsistency because the relative importance of future time periods does not change as time passes: under exponential discounting, the relative importance that I attach to tomorrow and the next day today

[12] On the inherent costliness of establishing precommitment devices, see McClennen (1990, pp. 196–98; and 1997, pp. 233–234) and DeHelian and McClennen (p. 326).

does not change once tomorrow arrives (p. 172).[13] Of course, both will be more important as they get closer, but they will have the same *relative* importance. In contrast, under hyperbolic discounting, as the future gets closer, the difference between upcoming periods grows larger, proportionally speaking, than when they are far off: their relative importance changes. Today, I deem tomorrow to be more important than the next day, and tomorrow, I'll not only judge that day and the following one to be more important than I do today (as I would even with exponential discounting), but I'll judge tomorrow's superiority in importance, relative to that of the next day, to have increased—so much so that I'll take option A tomorrow rather than, as I want today to do, waiting for option B.

So the blame for my pragmatic ineffectiveness may be shifted from my engaging in a form of straightforward maximization to my engaging in a dynamically inconsistent form of time discounting. But note that in order for straightforwardly maximizing theories to be defended against the charge of pragmatic ineffectiveness in this way, the defender of those theories must assume the additional burden of showing that such discounting is irrational. And he must do so in some *other* way than by claiming that the avoidance of such discounting is necessary in order for the agent to avoid pragmatically ineffective dynamic inconsistency, because this is plainly false: the agent can, instead, eschew straightforwardly maximizing choice in favor of indirectly maximizing choice.[14]

[13] In other words, the dynamic consistency of my preferences does not depend on the value of *r* or other arbitrary features of the example; the only analogues to the results of note 8 are negative ones. First, no matter what the values of options A and B are, there is no positive value of *r* such that on day 1 I will prefer B-on-day-3 to A-on-day-2 and yet on day 2 I will prefer A-on-day-2 to B-on-day-3. Second, no matter what *r* is (as long as it is positive), and no matter what the values of options A and B are, there is no number *t* such on day 1 I will prefer B-on-day-*t* and yet on day 2 I will prefer A-on-day-2 to B-on-day-*t*.

[14] Are there independent grounds for regarding hyperbolic discounting (or other dynamically inconsistent forms of time discounting) as necessarily irrational? Elster maintains that dynamically inconsistent preferences are irrational (1983, p. 7) and that policymakers may refuse to honor their constituents' preferences to the extent that doing so is necessary to smooth out the inconsistencies that hyperbolic discounting may generate (2000, p. 165). But he also admits that "many psychologists and behavioral economists argue that discounting is hyperbolic" in the case of most people (2000, p. 25) and that much of human behavior cannot be explained on the hypothesis that most people's time discounting is exponential (2000, p. 25, n. 73). Why does Elster indict the time discounting of most people as irrational? He gives two reasons. The first is that "an individual who discounts the future very heavily, with little ability to defer gratification, is unlikely to have a very good life" (2000, p. 25). But as he admits, this criticism is applicable to dynamically consistent as well as to dynamically inconsistent forms of discounting (2000, p. 26), so it is not a criticism of dynamically inconsistent discounting once the rational permissibility of discounting has been assumed (as it has been here, and as it is by Elster elsewhere, I as noted earlier, in subsection 2.2). Elster's second reason for criticizing the time discounting of most people is that "an individual who is subject to hyperbolic discounting is liable to time-inconsistency" (2000, p. 26). But this of course only invites us to ask again the question with which we started: Why should we regard dynamically inconsistent forms of time discounting as irrational?

4.7 Third, and finally, a defender of straightforwardly maximizing theories might claim that we have no atemporal perspective from which to say that my life, or the world as a whole, goes worse if, tomorrow, I choose to option A rather than declining it in order to obtain option B the next day. For all that can be said, about the plan that an agent regards as best at any give time, is that it makes his life or the world go best, as judged *from that point in time*. And it may be objected that we cannot assume the availability of *any* coherent notion of an agent's life's or the world's going better or worse overall, as opposed to just going better or worse from particular temporally indexed vantage points.[15]

But there are two problems with this reply. First, it is hardly a reply that a defender of a straightforwardly maximizing theory is in a position to make, since such a theory would itself require recourse to a notion of a life's or a world's going well—unless it were to take the form of something like Parfit's present-aim theory (1984, pp. 117–120), which falls outside the scope of this dissertation. Second, there are plausible proposals for combining an agent's temporally-indexed perspectives into an atemporal perspective,[16] and although the precise form that such a proposal ought to take may still be (and may long remain) a matter a some debate, it is hard to see how the very *idea* of such a proposal can be rejected.

4.8 This section has been devoted to exploring the self-defeat that may arise from the non-strategically induced dynamic inconsistency that straightforwardly maximizing theories may exhibit. We have seen that agents who employ such theories as their decision procedures may be led, by way of dynamically inconsistency, to be forced to settle for outcomes that are worse than other agents, such as indirectly maximizing agents, could obtain. And although this problem may tend to arise when the agent engages in dynamically inconsistent time discounting, or violates a principle of choice such as the independence principle, this fact does not save straightforwardly maximizing theories from the charge of self-defeat unless it can be shown that *those* features of the agent's decision-making, not his being a straightforward maximizer, are to be held liable for the inferior outcomes he obtains.

[15] This position may seem all the more plausible in the light of Arrow's theorem about the impossibility of satisfactorily aggregating distinct individuals' preference orderings into a social preference ordering. For it might be thought that a similar theorem could be proved about the impossibility of satisfactorily aggregating the distinct temporally indexed preference orderings of an agent into an atemporal preference ordering for that agent.

[16] See, for example, McClennen (1990, p. 212; and 1997, p. 220 and p. 241). Gauthier criticizes McClennen's proposal (1997b, pp. 17–19) and offers his own (1997b, pp. 20–23).

5 Dynamic inconsistency, part 2: strategically induced, individual

5.1 In the last section, we considered cases of non-strategically induced dynamic inconsistency. In this section and the next, we focus on cases of strategically induced dynamic inconsistency. These can be divided into two distinct types of dynamic inconsistency, corresponding to two distinct types of self-defeat: individual and collective. Individual self-defeat arises when just a single agent adopts the theory in question as his decision procedure; collective self-defeat arises when everyone in a group of agents, or nearly everyone in a group of agents, adopts the theory in question as his decision procedure. It is important to keep this distinction separate from the distinction between a theory of rationality and a theory of morality. The former distinction is orthogonal, not parallel, to the latter. A theory of rationality can exhibit not only individual self-defeat—if the agent is made worse off by his acceptance of it as a decision procedure—but also collective self-defeat—if agents are, considered from the point of view of individual, instrumental rationality, made worse off by the acceptance of it by everyone in their group than by their acceptance of some other theory. Similarly, a theory of morality can exhibit not only collective self-defeat—if worse outcomes result from everyone's acceptance of it that from everyone's acceptance of some other theory—but also individual self-defeat—if worse outcomes result from one agent's acceptance of it than from that agent's acceptance of some other theory. In each case, the term 'individual' or 'collective' refers only to the scope of the acceptance of the theory, and not to the scope of the consequences to be considered. The latter may also be individual or collective, but this is determined by the theory under consideration (typically individual for theories of rationality and collective for theories of morality), not by how widespread is the adoption being hypothesized.

Our focus in this section is on *individual* self-defeat: specifically, the individual self-defeat that arises from strategically induced dynamic inconsistency. This phenomenon occurs in variety of structurally different circumstances of interaction, which we shall consider one at a time. The following table shows the subsections in which we shall consider cases of the indicated kinds.

		egoism	act consequentialism
making promises		5.2	5.3
making threats	offensive	5.4	5.5
	defensive	5.6	5.7
ignoring threats		5.8	
winning toxin money		5.9	

Note that we shall consider both egoism and act consequentialism for a given kind of case before moving on to the next kind of case.

5.2 We begin our survey with the individual self-defeat of egoism (as it arises, of course, from strategically induced dynamic inconsistency). The phenomena in which we are interested, then, are ones in which an agent who subscribes to egoism—choosing in every case whatever act will further his ends the most—ultimately does not further his ends as much as he would if he subscribed to some alternative theory of rationality. We can see this in cases of several kinds.[17]

First, consider a case that Gauthier (1994, pp. 692–693; see also 1988, pp. 208–209) adapts from Hume, in which we are to suppose that you and I are farmers. I want you to help me with my harvest this week, and you want me to return the favor next week. Each of us would most like to get help without giving it, but each of us still prefers cooperative harvesting to independent harvesting. And because you in particular prefer cooperative harvesting to independent harvesting, you will be willing to help me this week if, but only if, I sincerely promise you that I will return the favor next week. (I am a terrible bluffer, so only a *sincere* promise will work.) But let us suppose, in addition, that I will have no incentive to help you next week: by then I will have gotten the help from you that I wanted (or will not have gotten it, and will not then be able to get it); I am not much moved (and will not be much moved next week, either) by moral considerations favoring reciprocation or keeping my word; and I will soon be retiring to Florida, never to cross paths again with you or with the neighbors to whom you will surely report my rude conduct if I fail to reciprocate.[18] The key point is that what will further my

[17] Most of my discussion of the individual pragmatic ineffectiveness of egoism—to be more specific, the bulk of subsections 5.2, 5.4, 5.9, and 5.10—is based on my "The Toxin and the Tyrant: Two Tests for Gauthier's Theory of Rationality."

[18] The stipulation of the absence of reputation effects is necessary because some of Gauthier's accounts of constrained maximization have been taken to involve "a covert appeal to reputation effects (which have the effect of

ends the most next week—what will be optimal for me—is to refuse to help you, whether you have helped me this week or not.

Now, if I am an egoist, then I have a problem. For I can already see, now, that I will definitely refuse to help you next week. And this foreknowledge that I have, of my own egoistic behavior, prevents me from sincerely promising you that I will help you. As a result, you will, quite reasonably, refuse to help me this week, and our cooperative venture will never even get off the ground. More generally, I can never be admitted to any cooperative venture in which I would be called on to contribute after benefiting, because I cannot give the sincere promise of contributing that is needed for people to admit me to such ventures in the first place.[19] I can be admitted to ventures in which my benefiting would be contingent on my contributing (such as when I would benefit after contributing, and my contribution can be verified), since then my contributing would be optimal: it would be a means to my benefiting. Of course, I would *like* to be able to give promises of the kind I am unable to give, even at the cost of following through, since I prefer cooperative harvesting and other cooperative outcomes to independent harvesting and other non-cooperative outcomes; indeed what is especially frustrating about such cases is that it is not my aim to exploit other participants in the venture—I would be happy just to get in and get my fair share of the fruits of cooperation, along with everyone else. But as an egoist, I am excluded from such ventures. Gauthier sums up my trouble succinctly: "his very way of choosing affects the situations in which he may expect to find himself. And the effects are to his disadvantage" (1984a, p. 263).

In this case, even though I follow egoism successfully, the consequences are worse than they would be if I were not an egoist, and so egoism exhibits individual self-defeat. Now it may appear that in such a case I do not follow egoism successfully, because it may appear that my failure to sincerely promise is a failure on my part to perform the act that would have the best consequences for me (since such a promise would result in your helping me, which would help me most of all). But successfully following egoism entails only performing the act *of those open*

changing preferences, and thus transforming [a situation of this sort] into another, more easily solved game)" (Hampton, p. 274). Whatever may have been inferable from Gauthier's earlier accounts of constrained maximization, it is crucial to understand that there is, in the interpretation of constrained maximization employed here, no appeal to reputation effects of the sort that would make my reciprocation next week anything other than a net loss to me.

[19] Another such venture is Parfit's desert-breakdown case (1984, p. 7). Parfit's remedy for the pragmatic ineffectiveness of egoism in this case is for the agent to be "*trustworthy*: disposed to keep [one's] promises even when doing so would be worse for [one]" (p. 7).

to me that has the consequences that are the best for me; it does not entail performing acts that, while not actually open to me, would be open to me if I were an agent of some other kind. The kind of agent I am—or, to be precise, the kind of agent I expect to be next week—constrains what I can do. In particular, it constrains what I can sincerely promise that I will do next week.[20] And so while I follow egoism successfully, the options available to me leave me worse off—even, to repeat, given my successful implementation of egoism—than I would be if I were not an egoist. In this way, as Hollis puts it, "self-interest, even enlightened self-interest, turns out to be self-defeating" (p. 17).

Can an agent of some other kind—such as an indirectly maximizing agent—do better? Consider a constrained maximizer. Unlike an egoist, for whom optimality of acts is everything—for whom acts deemed non-optimal are *ipso facto* ruled out—a constrained maximizer in this situation would realize that the best plan she can make is one calling for reciprocation—helping you if you help her, and not helping you if you do not help her—because that enables her to give you the sincere promise that results in getting your help; and when she proceeds to reciprocate, the result is cooperative harvesting, which she prefers to independent harvesting. Admittedly, her *best* outcome would be to get your help but then to refuse to reciprocate; but this is not an outcome that she can plan on: her getting your help is contingent on her giving you a sincere promise of reciprocation, which is contingent on her genuinely planning to help you if you help her; and she cannot both plan *to reciprocate*, as needed for the sincere assurance, and plan *not to reciprocate*, as needed for her best outcome. Seeing that reciprocation is part of the best plan she can make, she judges that reciprocation will be rational—and, more to the point, she expects that she will reach this same judgment next week, too. As a result, she can give you the sincere promise on which your helping her this week depends, and the mutual-benefit game plays out perfectly.

Now it may be objected that the constrained maximizer does not take *full* advantage of the situation in which she finds herself, because she voluntarily forgoes an opportunity to exploit you, by getting your help but not reciprocating; but that is beside the point. For the point is that the constrained maximizer's outcome of cooperative harvesting is preferable to the egoist's outcome of independent harvesting. As Gauthier writes, "Although the [constrained maximizer] refrains from making the most of her opportunities, yet she finds herself with opportunities that

[20] For a thorough discussion of this point, see Gauthier (1997a, pp. 29–30).

the egoist lacks and so may expect payoffs superior to those that he can attain" (1984a, p. 265).[21] And so in this case, at least, constrained maximization proves to be pragmatically superior to egoism.

5.3 Like egoism, act consequentialism may exhibit individual self-defeat due to the inability of agents who accept it as a decision procedure to make promises. Kydland and Prescott, in an influential paper whose main conclusions we shall consider shortly, offer a simple example by way of illustration. Suppose that a policymaker wants to encourage economic growth. One obvious strategy would be to offer patents to inventors: by protecting inventors from competition for some interval of time following the introduction of their inventions in the market, patents enable them to charge monopoly prices instead of competitive-market prices during that interval. But suppose that the policymaker is known to be an act consequentialist. Will his pledge to grant and enforce patents be trusted by inventors? Once an invention has been created, the principal incentive for granting and enforcing a patent has disappeared (though other incentives such as reputation effects—the desire of the policymaker to be seen by future potential inventors as one who keeps his word—may remain). So a policymaker who is known to be an act consequentialist might pledge to grant patents to inventors and to enforce those patents, but as Kydland and Prescott put it, "Given that resources have been allocated to inventive activity which resulted in a new product or process, the efficient policy [assuming the absence of reputation effects] is not to permit patent protection" (p. 457). Since potential inventors would be all too aware of the act-consequentialist policymaker's incentive to deny patent protection, they would decline to innovate (unless, of course, they thought that other incentives, such as reputation effects,[22] would be sufficient to ensure the policymaker's cooperation). In contrast, an indirectly maximizing policymaker such as a constrained maximizer or a resolute chooser would be able to make a credible announcement of a patent policy and obtain the better outcome. The parallel between the second Humean farmer (trying to assure the other farmer of his help next week) and the patent policymaker (trying to assure inventors of patent protection) should be

[21] In the passage from which this quotation is taken, Gauthier refers to the constrained maximizer as a "conditional cooperator." For terminological consistency with other parts of my discussion, I have replaced Gauthier's term as indicated.

[22] But note that the more valuable an invention is anticipated to be, the less likely it will be that other incentives such as reputation effects will be weighty enough to make patent protection optimal. So, even if many inventors believe their patents will be protected, those developing the most valuable inventions may not.

clear: in each case, straightforwardly maximizing choice (be it egoism or act consequentialism) is individually self-defeating.

Kydland and Prescott's central result, for which the foregoing example is merely preparatory, concerns a policymaker with some control over inflation and unemployment. In their article—tellingly entitled "Rules Rather than Discretion"—they argue that economic policymakers who are in a position to control inflation end up doing worse if they have the freedom to respond optimally to changing conditions than they would do if they were bound by rules. The freedom to respond to changing conditions takes the form of accepting, as their decision procedure, what is known in some circles as optimal control theory—which Kydland and Prescott describe as follows:

> Optimal control theory is a powerful and useful technique for analyzing dynamic systems. At each point in time, the decision selected is best, given the current situation and given that decisions will be similarly selected in the future. (p. 473)

What Kydland and Prescott are considering, then, is essentially the use of act consequentialism as a decision procedure.[23] And they claim that the acceptance of such a decision procedure by policymakers will be pragmatically ineffective:

> We find that a discretionary policy for which policymakers select the best action, given the current situation, will not typically result in the [best consequences]. Rather, by relying on some policy rules, economic performance can be improved. (pp. 473–474)

In other words, act consequentialism may be pragmatically ineffective in certain matters of economic policymaking.

Kydland and Prescott's argument presupposes two empirical claims. The first, which is known to economists under the guise of a graph known as the Phillips curve, states that, in the short run, policymakers face a tradeoff between inflation and unemployment: when inflation is higher than people expect, the economy expands, causing unemployment to drop; when inflation is lower than people expect, the economy contracts, and unemployment rises (Mankiw, pp. 318 and 332). What is the explanation for this tradeoff? There are several theories (Mankiw, pp. 290–299), but one of the simpler and more intuitive ones is this. When a firm encounters an unexpected increase in the price that buyers will pay for its product, it is faced with the following question: Are prices all over the economy being driven up, in which case there is a general rise in

[23] Note the import of the phrase 'given that decisions will be similarly selected in the future': it signals that the act-consequentialist policymaker being discussed is one who employs what Strotz calls "the strategy of consistent planning" (see subsection 4.5, above).

prices and no relative rise in the price of this product; or is the price of this product being driven up more than other prices in the economy are? If it's the latter—a relative rise in the price of the firm's product—then the firm stands to gain if it takes advantage of the favorable environment for its product by increasing output, which tends to involve hiring more workers. If it's the former—a general rise in prices, with no relative rise in the price of the firm's product—then the higher price of the firm's product is an illusion and the firm has no reason to deviate from its planned level of output. Any firm, though, tends to have imperfect information about the state of the economy; in particular, it doesn't know to what extent any rise in the price of its product is a relative rise in the price of its product, and to what extent it is part of a general rise in prices. Because the firm has no way of being sure that an increase in the price of its product is *entirely* due to a *general* increase in prices, the firm hedges its bets and increases output to some extent, hiring more workers as a result. Other firms react similarly to the unexpected inflation, and unemployment falls (Mankiw, pp. 296–297). Although there are several other explanations of the short-run tradeoff between expected inflation and unemployment—another one is summarized by Elster (2000, p. 151)—they all lead to this conclusion: "If the price level is higher than the expected price level, output exceeds its natural rate [and unemployment decreases]. If the price level is lower than the expected price level, output falls short of its natural rate [and unemployment increases]" (Mankiw, pp. 301–302). This fact is widely presupposed in monetary and fiscal policy. For example, it is presupposed by the U.S. Federal Reserve when it tries to keep inflation in check by tightening the money supply, since it is *by slowing growth* (which includes hiring) that tightening the money supply is supposed to keep inflation in check.

This tradeoff, as I said, is a short-run phenomenon, and the reason for this is the second of the empirical claims that I mentioned above: there is a natural unemployment rate that actual unemployment tends to approximate, but there is no natural inflation rate that actual inflation tends to approximate. There are several factors contributing to the existence of the natural unemployment rate. One is frictional unemployment: workers periodically lose their jobs and take some time to find new jobs; the time workers spend unemployed is increased by unemployment insurance, which makes finding a new job less urgent (Mankiw, pp. 121–125). A second explanation is wage rigidity: "the failure of wages to adjust until labor supply equals labor demand" (Mankiw, p. 126). This may be caused by any of several factors, including minimum-wage laws, union contracts that fix wages for certain periods of time, and the desire of employers to pay their employers more than they would earn in an economy of full employment, in order to reduce their temptation of workers to switch jobs or risk getting fired by slacking off.[24] So although unemployment surely varies with economic conditions, there are underlying factors that tend to make it hover around some level significantly above zero. In contrast, there is no natural inflation rate or natural gap between expected inflation and actual inflation. As Mankiw writes, "Eventually, expectations adapt. . . . In the long run . . . unemployment returns to its natural rate, and there is no tradeoff between inflation and unemployment" (p. 308).

The lack of a natural rate for either inflation or expected inflation makes lowering inflation a more natural long-term goal for a policymaker than lowering unemployment. And let us assume, for simplicity, that the policymaker has the freedom to make inflation as low as he chooses, by continuing to tighten the money supply until the economy slows down enough for inflation to fall to the level he desires. This, clearly, gives the policymaker some influence over what agents expect inflation to be.

To put the foregoing ideas into a convenient framework, let us represent the short-run Phillips curve with the equation $u = u^n - 2(\pi - \pi^e)$, where u is the unemployment rate, u^n is the natural unemployment rate, π^e is the expected inflation rate, and π is the (actual) inflation rate. That unemployment falls below its natural rate by approximately *two* percentage points for every percentage point by which inflation exceeds expected inflation is suggested by empirical data (Mankiw, p. 309).[25] And let us say that the natural unemployment rate is 6 percent, which is also in the ballpark suggested by the data (Mankiw, p. 119). Then the short-run Phillips curve becomes $u = 6 - 2(\pi - \pi^e)$. We will be considering cases in which the expected inflation rate is 0, 2.4, 4.32, 6, and 12, so let us go ahead and note the Phillips curve for those cases:

1. If $\pi^e = 0$, then $u = 6 - 2(\pi - 0)$, or $u = 6 - 2\pi$.

[24] Henry Ford appears to have had something like this in mind when explained his decision to pay his workers about twice the wage prevailing in the market by saying that "We wanted to pay these wages so that the business would be on a lasting foundation. . . . A low wage business is always insecure" (Mankiw, pp. 131–132).

[25] But, as I said earlier, unemployment returns to its natural rate. How fast? To answer this question, I should say that the general rule is not as simple as what I stated: that unemployment falls below its natural rate by approximately two percentage points for every percentage point by which inflation exceeds expected inflation. Rather, the first part of the rule is that unemployment falls below its natural rate by approximately two *percentage-point–years*, meaning that inflation falls below its natural rate by approximately two percentage points for one year and then returns to its natural rate, or falls below its natural rate by approximately one percentage point for two years and then returns to its natural rate, or falls below its natural rate by approximately half a percentage point for four years and then returns to its natural rate, and so on (Woodward, p. 170). For simplicity, I assume throughout that when unemployment falls below its natural rate by m percentage-point–years, it falls below its natural rate by m percentage points for one year and then returns to its natural rate.

2. If $\pi^e = 2.4$, then $u = 6 - 2(\pi - 2.4)$, or $u = 10.8 - 2\pi$.
3. If $\pi^e = 4.32$, then $u = 6 - 2(\pi - 4.32)$, or $u = 14.64 - 2\pi$.
4. If $\pi^e = 6$, then $u = 6 - 2(\pi - 6)$, or $u = 18 - 2\pi$.
5. If $\pi^e = 12$, then $u = 6 - 2(\pi - 12)$, or $u = 30 - 2\pi$.

Each of these five equations represents a line that, in turn, represents an infinite set of inflation-unemployment bundles among which the policymaker may choose, depending on which line describes the environment the policymaker faces—i.e., depending on what the expected inflation rate is. (In Figure II.2, in which only the first four are shown, they are the straight, downward-sloping lines.[26]) We said that the policymaker has the power to manipulate the money supply to make inflation go to whatever level he chooses. But once the public comes to expect a certain level of inflation, the Phillips curve for that level of inflation operates as a constraint on what inflation-unemployment bundle the policymaker can choose. So, for whatever level of inflation the policymaker chooses, the Phillips curve that he faces determines the resulting level of unemployment. Thus, if the public expects inflation to be 6 percent, the policymaker can give it a pleasant surprise of inflation of just 2 percent, but only at the cost of raising unemployment to 14 percent (since, as implied above, lower-than-expected inflation will cause firms to think that—to some extent, at least—there's less demand for their products, and so they'll lay off workers).

To complete the picture, we need to consider the policymaker's aims: what does it mean for him to optimize? Let us say that there is a *loss function* relating every inflation-unemployment bundle to a score that the policymaker tries to minimize, and let us say that the loss function is $\pi^2 + u^2$. This formulation captures the thought that inflation and unemployment are equally bad, as well as the thought that not just the absolute social costs of inflation and unemployment, but also the marginal social costs of these economic maladies, increase as their levels rise—hence the exponents.[27] This formulation also includes the nontrivial assumption that zero is the optimal level of inflation.[28] But this assumption is not unreasonable, and it certainly

[26] When drawn by its originator, A. W. Phillips, the Phillips curve was a curve, not a line. But these days it is drawn as a line, since "A linear Phillips curve fits the data extremely well" (Blinder, p. 19). For an account of how the modern Phillips curve differs from the original Phillips curve, see Mankiw (p. 304).

[27] That central bankers' objective functions tend to be quadratic in this way is confirmed by Blinder (p. 4).

[28] This assumption is nontrivial because "inflation is a tax on reserves and currency, and a more informed public might prefer some positive or negative inflation rate" (Kydland and Prescott, p. 480). There are at least two reasons why it might prefer a positive inflation rate. First, since real interest rates equal nominal interest rates minus inflation, it is easier to have extremely low real interest rates—which may sometimes be desirable—if inflation is positive, since it is hard to have extremely low nominal interest rates. Second, since real wage increases equal nominal wage increases minus inflation, it is easier to have flat or negative real wage increases—which may

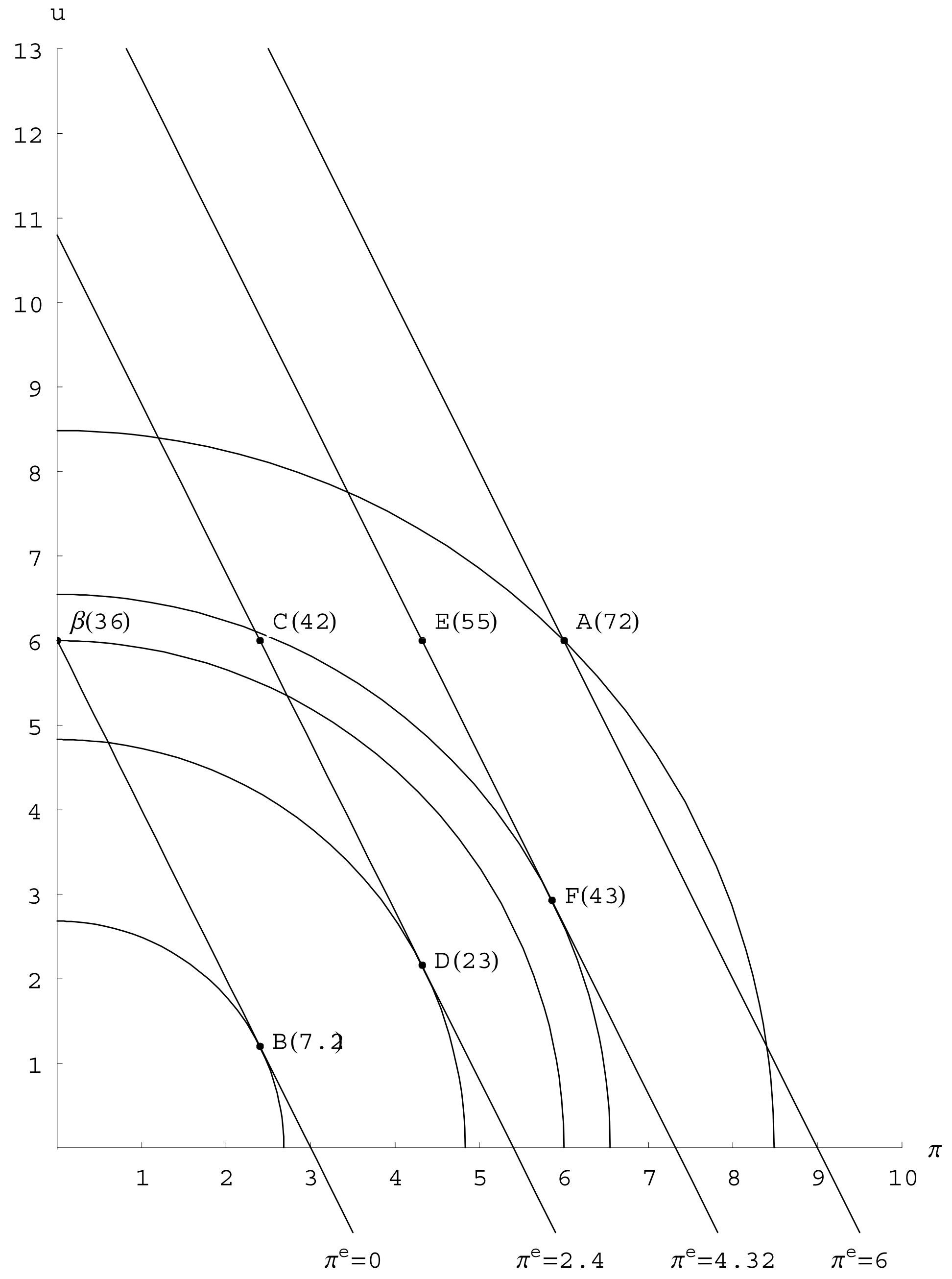

Figure II.2

simplifies the arithmetic. One final simplifying assumption implicit in this loss function is the absence of reputation effects, which we touched on in the patent case. With the loss function so specified, we can draw indifference curves, also shown in Figure II.2. Since these are

sometimes be desirable—if inflation is positive, since it tends to be nearly impossible to lower nominal wages. Positive inflation, then, gives policymakers and other major decision-makers flexibility that they would otherwise lack.

indifference curves for a *loss* function, the most preferable indifference curve is the one closest to the origin.

Let us consider a policymaker facing inflation and unemployment of 6 percent (point A in Figure II.2, with a loss value of $6^2 + 6^2$, or 72). What can he do? Let us suppose that he is an act-consequentialist policymaker—one who optimizes at every choice point—and that he starts by announcing his intention to lower inflation to 0 percent. If expected inflation then becomes 0 percent, he finds himself on the most favorable of the Phillips curves shown (the line farthest to the left), and he chooses the most desirable of the points on that line: point B, where inflation is 2.4 percent and unemployment is 1.2 percent, with a loss value of 7.2, which is clearly better than the loss value of 36 that would result from following through with his announced intention and settling for point β. As Romer explains, "this 'reneging' on the commitment raises social welfare" (p. 402). But if he is known to be an optimizer, then expected inflation won't really be 0; the public will anticipate his decision to choose point B and expected inflation will be 2.4 percent. Then the policymaker will be operating on the next Phillips curve—the one where π^e = 2.4—and if he sets inflation at 2.4 percent anyway, then unemployment will remain at 6 percent: point C, with a loss value of about 42. But the policymaker will *not* set inflation at 2.4 percent; he can do better, by taking advantage of *that* Phillips curve and choosing the most desirable of the points on that line, which is point D, where inflation is 4.32 percent and unemployment is 2.16 percent, with a loss value of about 23. But this, too, will be anticipated, and the policymaker will find himself on the even less-desirable Phillips curve corresponding to expected inflation of 4.32 percent (the third Phillips curve from the left). If (as assumed) he optimizes within this constraint, he'll reject the outcome corresponding to 6-percent unemployment (point E, with a loss value of about 55) in favor of point F, with a loss value of about 43. But this, obviously, is no more stable a stopping point—no more of an equilibrium outcome—than is B or D. Clearly we have to iterate these calculations and look for a limiting point.

It might be thought that iterating these calculations would show that the policymaker is driven back to the status quo of 6-percent inflation and 6-percent unemployment (point A). Then the policymaker would be the very model of impotence. But the policymaker's predicament is even worse. For if the policymaker were to settle on point A, then he would not be taking full advantage of the Phillips curve on which he finds himself. (To see this, notice that the line representing that Phillips curve—the line labeled π^e = 6—has points inside the indifference curve

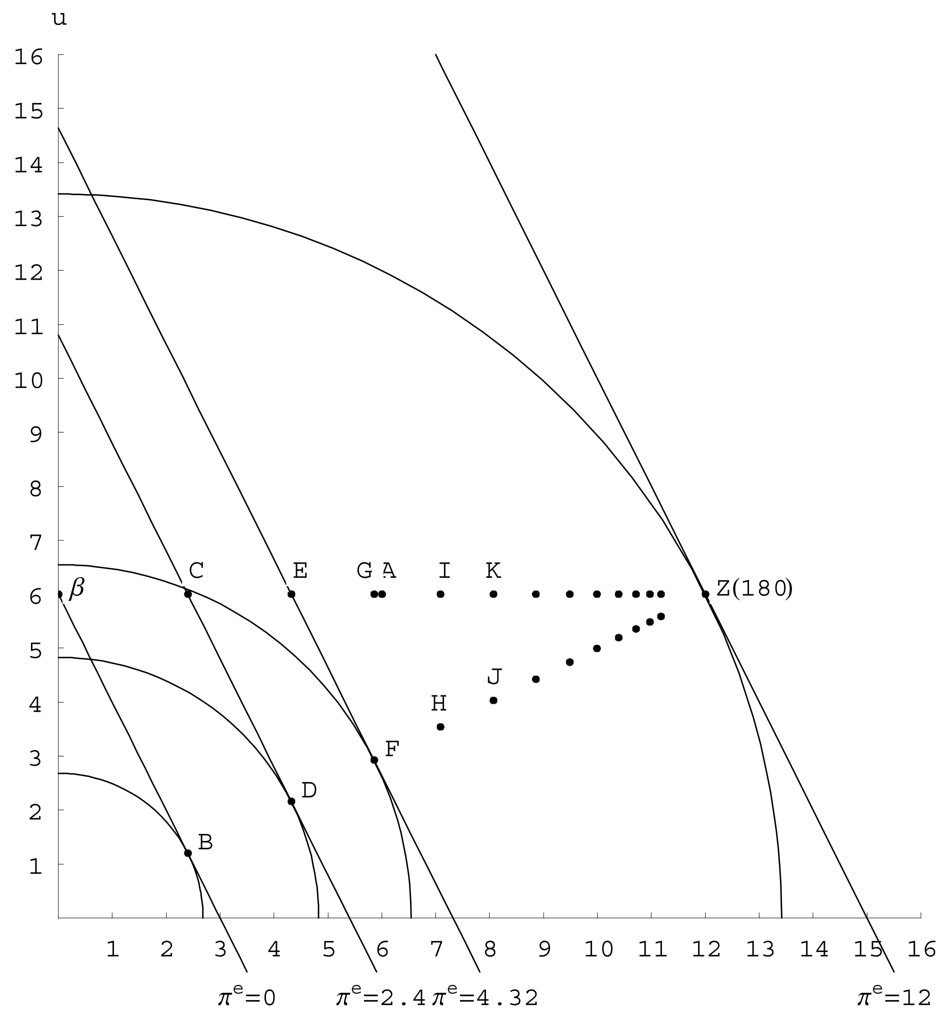

Figure II.3

corresponding to a loss value of 72. This means that point A is inferior to points where unemployment is lower in the same way that points C and E are inferior to certain alternatives to *them*.) So A is not an equilibrium outcome either. So where does a policymaker on the A–B–C–D–E trajectory end up? To discover this, we need to take a broader view of the sort found in Figure II.3. If we think of the policymaker as bouncing back and forth between the horizontal line (not drawn) on which A, C, and E lie and the sloped line (also not drawn) on which B, D, and F lie, then we can see that the policymaker bounces back and forth until the lines meet, at point Z, where unemployment is 6 percent but inflation is 12 percent (for a loss value of 180!). The policymaker, who we thought was the model of impotence, turns out to be something even worse: the very model of counter-productivity.

We have seen the policymaker's asymptotic approach to point Z. What is the rationale for Z's being the equilibrium outcome? Well, Z has a pair of traits that none of the other outcomes has. First, it is a point where an indifference curve is tangent to a Phillips curve. So, unlike points such as A, C, and E (and other points on the line they are on), it is a point from which the policymaker has no incentive to deviate. Second, it is a point where actual inflation is equal to expected inflation (12 percent). So, unlike points such as B, D, and F (and others points on the line they are on), it is a point that is genuinely an option for the policymaker, in the sense of being compatible with the public's rational expectations. In short, the public can see from Figure II.2 and Figure II.3 that if it were to expect inflation to be anything less than 12 percent, then the policymaker would find himself on a Phillips curve that would give him an incentive to set actual inflation at some rate higher than the expected rate, in order to exchange some of the expected decrease in inflation for a more valuable decrease in unemployment. So the public knows that if it were to expect inflation to be anything other than 12 percent, it would find its expectation falsified; as a result, the public settles on the expectation that inflation will be 12 percent. Then, the policymaker optimizes subject to this constraint and sets inflation at 12 percent. For these reasons, Kydland and Prescott conclude that "select[ing] that policy which is best, given the current situation . . . results in excessive rates of inflation without any reduction in unemployment" (p. 477).

But now let us suppose that our policymaker is resolute: once he announces a plan, he sticks to it, even if that involves forgoing some further benefits. He is committed to tightening the money supply as much as necessary in order to make inflation go to 0 percent. For this policymaker, the choice is easy: announce that inflation will be 0 percent. Assuming that the public knows that he is resolute, expected inflation will be 0 percent, and the result will be inflation of 0 percent and unemployment of 6 percent —point β, with a loss value of 36. Through his willingness to forgo the further gains sought by the act-consequentialist policymaker, the resolute policymaker ends up with a better result. As Blinder writes, "a central bank that regularly reaches for short-term gains will, on average, produce more inflation but no more employment than a central bank that is more resolute" (p. 39).

Admittedly, the example under discussion is not completely general: if the initial situation were one in which inflation is much higher than unemployment (to be precise: more than twice as high as inflation), then there would be room for the optimizing policymaker to lower inflation to some extent without being confronted with the temptation we have been discussing, and so the policymaker could credibly announce a policy of lowering inflation to that extent. The key, of course, would be for inflation to start out so high (relative to unemployment), and for the policymaker's desired lowering of inflation to be so modest, that inflation would *remain* as serious a problem as unemployment. But when either inflation is not initially that high or the policymaker wants to lower it more aggressively, then the temptation arises again and the policymaker is back in the predicament we have been discussing.

To compare the pragmatic effectiveness of act consequentialism and that of resolute choice, consider the following summary by Elster:

> For the government [i.e., the policymaker], the best outcome would be to make a credible announcement of zero inflation and then to proceed to inflate; the next best to make a credible announcement of zero inflation and then proceed not to inflate; and the worst outcome is a credible – because self-fulfilling – announcement of a positive rate of inflation. Because an announcement of zero inflation is never credible, the government is stuck with the worst outcome. (2000, p. 152)

Neither the resolute policymaker nor the act-consequentialist policymaker can be expected to achieve the best outcome: that obtained by making a credible announcement of zero inflation and then proceeding to inflate. The resolute policymaker would not take the second step—proceed to inflate—while the act-consequentialist policymaker could not take the first step—make a credible announcement of zero inflation. But the resolute policymaker can be expected to achieve the second-best outcome (announce and implement zero inflation), while the act-consequentialist policymaker can be expected to end up with the worst outcome (announce and implement positive inflation). Thus the self-defeat of act-consequentialist policymaking in these circumstances.[29]

[29] The foregoing discussion suggests that one popular explanation for the independence that many central banks have from politicians is only part of the story. It is often claimed that central banks should be independent from political influences in order to keep them from being used for short-term political purposes. The thought, as Elster puts is, is that "If the government has direct control over monetary policy, it may use that instrument to enhance its chances of reelection rather than to serve the public interest" (2000, p. 153). But Elster points out that the foregoing analysis shows that "an independent central bank may be required even if the government is motivated solely by considerations of the public interest" (Elster 2000, p. 153). The problem with the act-consequentialist policymaker, after all, is not that he is selfish or shortsighted—not, as Kydland and Prescott put it, that he is "stupid or evil" (p. 487). Indeed his being well-informed and well-intentioned are crucial premises in the argument showing his ineffectiveness. The problem is simply that he optimizes with society's best long-run interests at heart.

An interesting implication is that it is not necessary for a policymaker not to be an optimizer in order for him to avoid pragmatic ineffectiveness; an optimizing policymaker may be as effective as a resolute one if his loss function is different. For example, if he is especially inflation-averse—more averse to inflation than he would be if

5.4 Making a sincere *promise* to perform non-optimal conduct, even when giving such a promise would itself be optimal, is not the only potentially optimal conduct that a straightforwardly maximizing agent is unable to perform. A central problem in recent work on deterrence is that such agents cannot make sincere *threats* to perform non-optimal conduct.[30] Borrowing another example from Gauthier, let us suppose that I buy widgets from you in large quantities (1996, pp. 224-25). I know that what I pay you far exceeds what you need to be paid in order for our transactions to be worth your while, and I know that you would give in to my demand for a discount if you feared that I would take my business elsewhere. So you would give me the discount if, but only if, I sincerely threaten to take my business elsewhere. (Again, I am a terrible bluffer.) But, in addition, taking my business elsewhere will be non-optimal for me: even if you do not give me the discount, the price you charge for widgets will remain, as it is now, the lowest price I can find, and there will be no other consequences to worry about. In short, what will be optimal for me is to continue to buy widgets from you, whether you have given me the discount or not.

If I am an egoist, then I have the same problem as before: I cannot sincerely tell you now that what I do second will depend on what you do first, since I know that is not true. And so you will not give me the discount that I want.[31] Admittedly, in this case my motives are harder to sympathize with, because instead of tying to participate in a venture that helps us both, I am

his priorities mirrored those of the public—then he may obtain better outcomes than if he is not "overly" averse to inflation. As Mankiw points out, "Perhaps this is why even liberal politicians who are more concerned about unemployment than inflation often appoint conservative central bankers who are more concerned about inflation" (p. 341; see also p. 332; Romer, pp. 407–409; and Blinder, pp. 46–48). So a policymaker needn't avoid being an optimizer in order to avoid pragmatic ineffectiveness; he may also accomplish this by optimizing with respect to goals that diverge from those of the public he serves. The public, then, may maximally advance its interests by appointing, as policymakers, individuals who explicitly repudiate its conception of what consequences are best.

[30] See Gauthier (1984b, pp. 298–302) and Kenny (pp. 21–22 and pp. 37–38). To understand the nature of threats as they are conceived of here, note Parfit's distinction between warnings and threats: "When I say that I shall do X unless you do Y, call this a warning if my doing X would be worse for you but not for me, and a threat if my doing X would be worse for both of us" (1984, p. 20). The crucial point is that an agent *threatens* to perform an act only if the act is non-optimal for the agent: otherwise, the agent only *warns* that he will (under certain circumstances) perform the act. Admittedly, threats and warnings may be distinguished differently; Elster says that what distinguishes threats from warnings is that it is up to the agent who issues the threat to decide to follow through on it or not, and that this is not true of warnings (2000, p. 38). But conceiving of threats in this way does not isolate the class of cases that interest us: although all of the cases that interest us are Elsterian threat cases, not all Elsterian threat cases interest us.

[31] In discussing the assurance case I said I would like to be able to give certain assurances "even at the cost of following through." Can we make an analogously strong claim about threats? No, because the threats I would like to be able to make are threats without follow-through costs, since we supposed that, if made, they would be successful: we supposed that "you would give me the discount if . . . I sincerely threaten to take my business elsewhere."

trying to exploit you, and we might be glad that I cannot. But however socially desirable this outcome may be, it is still a failure of instrumental rationality: it is a failure to use the means at my disposal—in this case my strong bargaining position—for the maximal furtherance of my ends. In a sense, only my thoughts (my thoughts about what it will be rational for me to do, if you ignore my threat) stand in my way. It is as if I have the opportunity to pick your pocket, but am somehow handcuffed from the inside.

But how would our constrained maximizer fare in my place? She would see that the best plan she can make is one calling on her to take her business elsewhere if you do not capitulate, because if she makes this plan, then she can issue the sincere threat that, by hypothesis, results in her best outcome. Of course, this would be the best plan for an egoist to make, too, if he were capable of making it; but because it would call on him to perform a non-optimal act—or, at least, to be prepared to perform a non-optimal act, depending on whether you capitulate—he cannot adopt it. Because a constrained maximizer can plan to perform acts that are not optimal (if they are called for by plans that are optimal, as this one is), she can issue the sincere threat that results in your capitulation, outdoing the egoist yet again.

5.5 The constrained maximizer outdoes the egoist because she is what Parfit calls a *threat-fulfiller*: an agent who always follows though on her threats (1984, p. 20). Do threat-fulfillers outdo act consequentialists, too? Initially it may appear unlikely that they would, since the keeping a threat involves harm to *both* parties, and while this sort of outcome may be understandable in the case of egoism—since egoism tells the agent to advance his own aims, regardless of whether he must exploit others in order to do so—it may seem out of place in the context of act consequentialism, since this theory tells the agent to advance, in effect, everyone's aims. It turns out, though, that the reasoning from subsection 5.4 applies even when we replace egoism with act consequentialism.

To see this, suppose that the benefit our agent seeks to obtain from the other agent is not just a benefit from the point of view of egoism, but a benefit from the point of view of act consequentialism. And suppose that her threat is to harm the other person not at her own expense, but at the expense of the public good (which must be the case in order for the agent to be issuing a *threat* in the context of act consequentialism). Then our agent's being a threat-fulfiller enables her to credibly issue the threat, which is something that she would be unable to do if she were an act consequentialist.

Plausible examples are hard to come by, but imaginary ones are easy to invent. Suppose that there is a mad scientist working in his basement, miraculously coming up with cures for all sorts of diseases but refusing to make any of those cures available to the public. Let us say that the only way the scientist can be induced to turn over his work is by way of a credible threat to destroy him and his lab. Since this permanent destruction of the scientist's results would be worse for the public than, e.g., waiting for him to die (at which point the public could just take his results), the threat would not be credible if issued by an act consequentialist policymaker. But if the policymaker is a threat fulfiller (and knows herself to be one), then she could sincerely issue the threat, bringing about the best outcome (best, of course, from the act-consequentialist point of view, though not from the scientist's point of view).[32]

5.6 We may regard the threats discussed in subsections 5.4 and 5.5 as *offensive* threats, since their purpose is to enable the agent to gain some advantage that she did not previously have.[33] That is, their purpose is to enable the agent to advance her aims (whether they be egoistic or act consequentialist). There are also *defensive* threats, whose purpose is to enable the agent to avoid losing ground in terms of her aims.

As I suggested at the beginning of subsection 5.4, cases of deterrence may be the most widely discussed cases of defensive threats, and of threats in general. If an agent known to be an egoist fears attack from an enemy, and the only possibility of deterring the attack is for him to credibly threaten to set off an explosive that will kill both him and his attacker, then he will be unable to deter the attack, since egoists do not follow through on threats. As a result, the agent will be subject to attack. In contrast, a non-egoistic agent who meets Parfit's criterion for being a threat fulfiller would be able to issue the threat and, assuming her would-be attacker backs down, would then be spared the prospect of having to either follow through on it or have her threat shown to be empty. Comparing the relative prospects of the egoist and the threat fulfiller in threat cases, then, further illustrates the self-defeat of egoism (Parfit 1984, p. 20).

5.7 The parallel in the case of act consequentialism is obvious, and is illustrated in the following example. If two countries—A and B—are on the verge of war, such that the only

[32] Here is another example: hijackers take over a plane. An act-consequentialist official knows that the hijackers prefer (1) money and freedom to (2) being captured and getting no money, and prefer that to (3) being killed. The act-consequentialist official would like to be able to credibly threaten to blow up the entire plane—hijackers, passengers, and all—if they don't surrender. But he can't, of course, if (as one would expect) he would regard that outcome as worse than negotiating with the hijackers.

[33] Even in the hijacking case, the threat is an offensive one, since the policymaker is trying to bring about some improvement in the (admittedly dire) status quo, not simply to maintain it.

deterrent available to the leader of A is to credibly threaten to retaliate in a way that will not only harm B but also produce worse consequences overall, then the leader of A had better not be known to accept act consequentialism as his decision procedure. For if he is, then the leader of B will know that the leader of A will not retaliate if B strikes first, and A will be defenseless. In contrast, if the leader of A were known to be a threat fulfiller, then he could credibly threaten retaliation. Then the leader of B would know that any first strike on his country's part would be met with retaliation, and he would back down. (Let us say, to simplify matters, that the leader of B is an egoist, not an act consequentialist.) So if the leader of A is an act consequentialist, his options are restricted: he cannot issue a credible threat of retaliation. This makes him less effective, as an agent of bringing about good consequences, than he would be if he were an agent of a certain other kind. In effect, the leader's problem is that the intention of retaliation that he would like to be able to form is an intention to perform an act that he regards as immoral, and he can form this intention only by repudiating some of his moral views. As Kavka writes, "He is a captive of the prison of his own virtue, able to form the requisite intention only by bending the bars of his cell out of shape" (1978, p. 291). The act-consequentialist agent, no less than the egoistic agent, is prevented from using threats to obtain advantages that his bargaining position makes available to him.[34]

Another example can be found in Kydland and Prescott's paper, the central results of which we have already examined in the context of promises. This example involves flood-control measures:

> suppose the socially desirable outcome is not to have houses built in a particular flood plain but, given that they are there, to take certain costly flood-control measures. If the government's policy were not to build the dams and levees needed for flood protection and agents knew this was the case, even if houses were built there, rational agents would not live in the flood plains. But the rational agent knows that, if he and others build houses there, the government will take the necessary flood-control measures. Consequently, in the absence of a law prohibiting the construction of houses in the flood plain, houses are built there, and the army corps of engineers subsequently builds the dams and levees. (p. 477)

[34] Nixon's chief of staff, H. R. Haldeman, reports that Nixon had a theory about how to get the North Vietnamese to believe that Nixon would use "excessive force" and hence to capitulate. He quotes Nixon as saying in a private conversation, "I call it the Madman Theory, Bob. I want the North Vietnamese to believe I've reached the point where I might do *anything* to stop the war. We'll just slip the word to them that, 'for God's sake, you know Nixon is obsessed about Communism. We can't restrain him when he's angry—and he has his hand on the nuclear button'—and Ho Chi Minh himself will be in Paris in two days begging for peace" (p. 83).

In this situation, act-consequentialist policymakers could be exploited by agents who know that if they build houses in flood plains, the policymakers will protect them; whereas policymakers capable of threatening that they will not protect agents—and it will be a threat, not just a warning, since (by hypothesis) neglecting them will be a sacrifice in terms of overall good, once they have built houses in the flood plain—can bring about the superior outcome of no (socially costly) construction in flood plains.[35]

There are, of course, cases of other kinds as well. One that has received considerable attention, but that we shall not examine here, is that of punishment: whether an act-consequentialist official can credibly threaten to inflict on someone a punishment that he believes will have no benefits in terms of deterrence, compensation for victims, or other forward-looking factors.[36] But for our purposes, the cases already discussed suffice to show how threat cases display the self-defeat of act consequentialism.

5.8 We have been discussing the inability of straightforwardly maximizing agents to fulfill—and, hence, even to sincerely issue—threats. But this is not the only problem such agents have with threats. To see this, imagine a character whom Parfit calls a threat ignorer: an agent who "always ignores threats, even when he knows that doing so will be worse for him" (1984, p. 20). And suppose that there are two agents, one an egoist and the other a threat ignorer, to whom a third agent—herself a threat fulfiller—is considering making an offensive threat. The threat fulfiller will realize that if she makes the threat to the egoist, then the egoist will capitulate, and so she has every incentive to make the threat and exploit the egoist to no end. The threat fulfiller will also realize, however, that she has no incentive to threaten the threat ignorer, because she will only end up having to follow through on her threat after failing to secure the desired capitulation from the threat ignorer. The threat ignorer, then, is immune to the bullying that the egoist can expect to encounter; he is insulated from attacks to which the egoist is vulnerable. He has superior defensive capabilities.

Does the threat ignorer have superior offensive capabilities, too? It would seem not, since being a threat ignorer essentially amounts to being a threat fulfiller in a certain class of cases:

[35] John McPhee quotes one observer of this phenomenon as saying, "There are a lot of disasters waiting to happen. The people want to live in these areas. When they buy houses, they don't know what they are getting into. The entire county ends up paying for these people's problems. The people should be assessed for these things. They are localized problems. The whole county is subsidizing people on the front line" (p. 248; see also p. 262).

[36] See, for example, Hodgson (pp. 83–110), Regan (pp. 66–82), and Kavka (1986, pp. 102–103 and pp. 151–152).

those cases in which the non-optimal behavior that one is prepared to engage in, if push comes to shove, is the behavior of not giving in to a threat. And since the agent is simply trying to maintain the status quo, not to improve upon it, he is engaging in characteristically defensive threat behavior. So the function of being a threat ignorer (i.e., of issuing that special class of threats just identified) is to avoid being exploited by others, not to exploit them oneself.

So threat ignorers do better than egoists defensively, and do no worse offensively. Like threat fulfillers, then, threat ignorers do better than egoists. The foregoing reasoning, along with the results of our earlier considerations of promises, leads to the conclusion, aptly summed up by Parfit, that the agent who enjoys the most pragmatic effectiveness is "a trustworthy, threat-fulfilling, threat-ignorer" (1984, p. 457). Such an agent, obviously, repudiates egoism in several distinct ways. Similarly, an agent can expect to bring about much better consequences overall if she repudiates act consequentialism in these same ways.

5.9 In both promise cases and threat cases, what being a straightforwardly maximizing agent prevents an agent from doing is forming certain intentions: forming intentions to perform acts that he knows will be non-optimal, such as contributing to a cooperative enterprise when he has already benefited, or following through on a failed threat. To focus more directly on the straightforwardly maximizing agent's inability to intend to perform non-optimal acts, consider Kavka's much-discussed toxin puzzle. An eccentric billionaire (without whom many a philosophical thought experiment would be utterly insolvent) offers me the following deal:

> He places before me a vial of toxin that, if I drink it, will make me painfully ill for a day, but will not threaten my life or have any lasting effects. . . . The billionaire will pay me one million dollars tomorrow morning if, at midnight tonight, I intend to drink the toxin tomorrow afternoon. He emphasizes that I need not drink the toxin to receive the money; in fact, the money will already be in my bank account hours before the time for drinking at arrives, if I succeed. . . . All I have to do is intend at midnight tonight to drink the stuff tomorrow afternoon. I am perfectly free to change my mind after receiving the money and not drink the toxin. (Kavka 1983, pp. 33–34)[37]

Now of course I am a lot more interested in getting that million dollars than I am in avoiding a day's illness. So although I would love to get the million dollars without drinking the toxin, I would still much prefer getting the million dollars—even along with drinking the toxin—to any

[37] For continuity with previous cases, I have replaced Kavka's second-person pronouns with first-person pronouns.

outcome in which I do not get the money. The money matters most. And all I have to do to get the money is form the intention to drink the toxin. But actually drinking the toxin will be non-optimal: when it is time to drink, I will either have the money or not; and the only thing left for me to decide will be whether to drink and thus make myself painfully ill. Naturally, I would prefer not to.

You know what is coming next. If I am a straightforwardly maximizing agent (whether an egoist or an act consequentialist), then I am out of luck: I can see now, and will continue to see at midnight tonight, that I will decline tomorrow afternoon to drink the toxin. This keeps me from forming the intention to drink, and thus from getting the million dollars. Obviously, if the order of events were reversed, so that my getting the money were contingent on my actually drinking, then I would be all set: for then drinking would be optimal, and I would gladly do it. Bu the situation is not like this. It calls for me to receive the benefit first and pay the cost second. Paying the cost will be non-optimal, and because I know I do not knowingly perform non-optimal acts, I cannot form the million-dollar intention. As before, what stops me from taking full advantage of the resources and opportunities at my disposal, or doing what will further my aims the most, is just my knowledge that this is the kind of agent I am: an agent who always chooses the act that furthers his aims the most. When the eccentric billionaire makes his offer, I cannot plan to take advantage of it; the best plan I can make is to try to just forget about the money. “At least I’ll have my health,” I’ll console myself.

But the constrained maximizer’s approach should be clear: she sees that the best plan she can make is to drink the toxin tomorrow afternoon, because that will result in her intending, at midnight tonight, to drink the toxin; and even when she drinks the toxin, at least she has secured the money. (Remember, the money matters most.) As in the farming case, the outcome she attains is not her best one, or *the* best one, which is getting the money and then not drinking, because this outcome is not something she can *plan* to attain—for if she plans on not drinking, then she will not intend to drink, and she will not get the money. Seeing, then, that drinking is called for by the best plan she can make, she judges that drinking will be rational or right, and she expects that she will reach this same judgment when it is time to drink. As a result, she intends to drink, earning the money that the straightforwardly maximizing is denied.

5.10 Hollow assurances, idle threats, and second-rate plans: these seem to be the hallmarks of the straightforwardly maximizing agent. Note that the straightforwardly maximizing agent's problem is not that he fails to take account of all of his options, or fails to foresee some of the consequences of some of those options, or fails to see how some of those consequences will affect the furtherance of his aims; we are considering an idealized straightforwardly maximizer agent, one who acts with perfect information about the world around him. His problem is deeper. His problem is that his very way of choosing—the mere fact that he is a straightforwardly maximizing agent, along with his awareness of this fact—prevents him from intending to perform non-optimal acts. And this prevents him from securing the benefits that certain situations offer.

But because an indirectly maximizing agent such as a constrained maximizer takes herself to have reasons to keep promises meeting certain criteria even though the required conduct is non-optimal, she can make sincere promises. Similarly, such an agent, by taking herself to have reasons to choose to follow through on (failed) threats meeting certain criteria (even though the required conduct is non-optimal), can issue sincere threats. And we have just seen how she takes advantage of the toxin case. In each of the three situations, then, an indirectly maximizing agent can adopt plans and intentions that a straightforwardly maximizing agent cannot, enabling her to avoid the inconveniences of straightforwardly maximizing choice and to secure the benefits that we would expect from *truly* maximizing choice. Thus, merely by limiting the range of things that an agent can intend to do, subscribing to egoism or act consequentialism can keep an agent from maximally furthering the ends his theory identifies as worth furthering. As a result, straightforwardly maximizing theories can be said to have many sources of individual self-defeat.

6 Interlude: dynamic inconsistency and autonomous effects

6.1 Before proceeding, it may be worthwhile to pause to consider, briefly but from a slightly more abstract perspective, the results of the previous section. For the principal implication—that the person who advances his aims the most is one who can intend to perform non-optimal acts—is plainly counter-intuitive, at least from a certain perspective. For if a person could intend to perform non-optimal acts, then he might occasionally perform non-optimal acts, thereby frustrating his ends rather than furthering them. And it might seem that a person is bound to end up worse off if he has this non-optimal conduct flaring up from time to time. We have seen how, in specific cases, an indirectly maximizing agent can expect to fare better than a straightforwardly maximizing one. But is there a more general explanation?

The fact that this question even arises alerts us to a distinction that should be made explicit at this point: the distinction between (1) an agent who optimizes in each of his acts and (2) an agent who optimizes on the whole—where "optimizes on the whole" means "chooses with the effect of doing as well, on the whole, as his resources permit." An agent of the first type might be an egoist or an act consequentialist, but in either case the crucial fact about him is that he optimizes at every choice point. And the results of the previous section suggest that, in being an agent of the first type, he is very likely not an agent of the second type. Thus the truth of Kavka's remark that "it may sometimes promote one's aims *not* to be disposed to act to promote one's aims" (1978, p. 293).[38] We can, then, rephrase our question in these terms: How does it happen that agents who optimize in each of their acts fail to optimize on the whole?

6.2 The answer, of course, is not that straightforwardly maximizing agents fail to make the most of their opportunities they are presented with—for we have abstracted from problems of implementation and desires that may be incompatible with making the most of one's opportunities—but that agents of certain other sorts are presented with better opportunities; and these better opportunities are *so much better* that these other agents do even better than straightforwardly maximizing agents do, even though they do not always make the most of their opportunities.

Several writers have made this point. Hodgson sums up the phenomenon clearly in his observation that

> There are alternatives open to those who accept personal rules approximating to the conventional moral rules of their society, but which are not open to those whose only personal rule is the act-utilitarian principle. Even though the former persons might not always choose, from the acts open to them, those with the best consequences, nevertheless the consequences of the acts which they do choose might be better than the best consequences which the latter persons could bring about through the alternatives open to them. (pp. 58–59)

Hodgson goes on to make similar remarks about egoism (p. 61), as does Parfit: he writes that if he is an egoist, then "The bad effects come, not from what I do, but from my disposition" (1984,

[38] A related claim is Adams's observation that "the motivational pattern that leads to more useful actions is not necessarily the more useful of two motivational patterns, on the whole" (p. 470).

p. 5). In contrast, Gauthier writes, “A constrained maximizer . . . benefits from her disposition, not in the choices she makes, but in her opportunities to choose” (1986, p. 183).

Romer uses similar language in regard to the policymaker who aspires to reduce inflation:

> If the policymaker announces that inflation will equal [the ideal rate] and the public forms its expectations accordingly, the policymaker will deviate from the policy once expectations are formed. The public’s knowledge that the policymaker would do this causes it to expect inflation greater than [the ideal rate]; this expected inflation worsens the menu of choices that the policymaker faces. (p. 402)

The key phrase, of course, is “menu of choices.” The act consequentialist does worse not because he fails to optimize once his menu of choices is set, but because his being an act consequentialist worsens the menu of choices that he faces.

6.3 The key to seeing how agents of different sorts can face different menus of choices is seeing that the intentions that an agent has can have effects other than those of the acts that are intended. To see this, recall the deterrence case. There, the intention to retaliate (conditional on a first strike by the enemy) has effects other than the (conditionally) intended retaliation: namely, the enemy’s capitulation. As Kavka points out, “such ‘evil’ intentions may pave the road to heaven, by preventing serious offenses and by doing so without actually harming anyone” (1978, p. 285). Kavka goes on to offer the following explanation:

> typically, the only significant effects of intentions are the acts of the agent (and the consequences of these acts) which flow from these intentions. However, in certain cases, intentions may have *autonomous effects* that are independent of the intended act’s actually being performed. In particular, intentions to act may influence the conduct of other agents. (1978, p. 291)

The having of an intention, then, can have effects that are independent of those of the performance of the intended act. And in the cases discussed earlier, the benefits enjoyed by a constrained maximizer—your cooperation in the assurance case, your capitulation in the threat case, the million dollars in the toxin case—are all effects of this kind. They flow from the constrained maximizer’s intentions, not from the acts intended.

Once individual self-defeat is understood in thee general terms, it is clear that the point extends more broadly than just to the theories we have been focusing on (egoism and act consequentialism). As Williams puts it, this point “is not just a point about the Utilitarian consciousness, although it very strongly applies to that: it is a point about the rationality of deliberation altogether" (1982, p. 163). Paraphrasing Sidgwick (p. 345), Williams writes that these considerations suggest that "the dictates of reason ought always to be obeyed, but it does not follow that the dictation of reason is always a good" (1982, p. 163).

7 Dynamic inconsistency, part 3: strategically induced, collective

7.1 In sections 5 and 6, we considered individual self-defeat: the self-defeat that a normative theory can exhibit when a single individual accepts it as his or her decision procedure for questions of the sort to which that theory pertains—whether they be questions of instrumental rationality (as in the case of egoism) or questions of morality (as in the case of act consequentialism). In this section, we shall consider collective self-defeat: the self-defeat that a normative theory can exhibit when all or nearly all of the members of some group accept it as their decision procedure for questions of the sort to which that theory pertains. We shall say that a theory exhibits collective self-defeat when there is some other theory whose acceptance (as a decision procedure) by everyone in some group would cause their aims to be better achieved than they would be by the acceptance of the theory under consideration.

Before proceeding it should be made explicit that in moving from the case of the individual acceptance of a normative theory as a decision procedure to the case of the collective acceptance of a normative theory as a decision procedure, we are not moving from the case of individual decision-making to the case of collective decision-making. We still suppose that decisions are made by individuals deciding (for themselves) what to do in the situations that they face, and any consequences resulting from the decisions and acts of some group may be regarded as consequences resulting from the decisions and acts of the individuals in that group. All that is meant by the collective acceptance of a normative theory as a decision procedure is acceptance by every individual for use in the situations that he or she faces. On the basis of this clarification it may appear that 'universal' would be a more appropriate term than 'collective', but having offered this clarification I shall proceed with the terminology already introduced.

In our examination of the individual self-defeat of egoism and act consequentialism in sections 5 and 6, we focused on a series of kinds of cases—promises, making offensive and defensive threats, ignoring threats, and winning toxin money—and considered egoism and act consequentialism together within our discussion of each kind of case. But in this section we shall first consider egoism in a variety of cases, deferring (until the latter part of this section) a considering of act consequentialism. For we shall see, when we consider act consequentialism, that the parallels that exist between it and egoism in the case of individual self-defeat (at it arises from strategically induced dynamic inconsistency) do not extend to the case of collective self-defeat (of the same sort).

7.2 We saw earlier (in subsections 5.2, 5.4, 5.6, and 5.8) the *individual* self-defeat of egoism in cases of promises and threats. Such cases also display the *collective* self-defeat of egoism. We begin by considering promises.

The kind of case to be discussed here, though only one of many in our elaborate taxonomy of kinds of cases in which straightforwardly maximizing theories are self-defeating, is of special interest in moral philosophy. For cases of this kind starkly pose the question of whether morality is consistent with rational self-interest, which—though not the focus of this dissertation—is widely discussed. Readers familiar with any of this literature will immediately recognize that the cases under consideration in this subsection are, essentially, versions of the prisoner's dilemma.

The key features of the prisoner's dilemma are conveyed in David Velleman's admirably pithy account of the scenario from which it gets its name:

> two people—say, you and I—are arrested on suspicion of having committed a crime together. The police separate us for interrogation and offer us similar plea bargains: if either gives evidence against the other, his sentence (whatever it otherwise would have been) will be shortened by one year, and the other's sentence will be lengthened by two. In light of the expected benefits, maximizing rationality instructs each of us to give evidence against the other. The unfortunate result is that each sees his sentence shortened by one year in payment for his own testimony, but lengthened by two because of the other's testimony; and so we both spend one more year in jail than we would have if both had kept silent. (p. 222)

Such a case plainly displays the collective self-defeat of egoism. For if we are both egoists, then we do not cooperate, and we each end up with a longer sentence than he needs to. We each end up with one more year in jail than necessary.

Although the prisoner's dilemma involves only two agents, it can easily be extended to an indefinitely large number of agents. A particularly clear example is provided by Elster:

> One hundred peasants all own land adjacent to a river. On each plot there are some trees and some land for cultivation. As the peasant families get larger, they decide to cut down the trees to get more land for cultivation. When the trees are

> cut, the roots lose their grip on the subsoil and land is lost to the river through erosion – not only the land on which the trees used to grow, but also some of the land previously used for cultivation. . . . I assume that trees being cut down on all adjacent plots is a necessary *and sufficient* condition for erosion on the individual plot. I assume, moreover, that if erosion occurs, any trees left on the plot will be lost to the river; finally I stipulate that the trees can provide wood for some useful purpose. It is then clear that each family will have an incentive to cut its trees, whatever (it believes that) the other families will do, for if the neighbors do not cut all their trees, the family can do so and get more land for cultivation, and if they do, it should cut its losses and at least get the wood. . . . As a game – it is, in fact, a Prisoner's Dilemma – the situation is trivial, since the solution is made up of dominant strategies. (1983, pp. 27–28)

Parfit suggests that we call such cases "Each-We Dilemmas" (1984, p. 91), since in such cases what *each* of us ought to do for himself conflicts with what *we* ought to do for ourselves.

What these dilemmas show is that the universal acceptance of egoism can have unfortunate consequences. Of course, the obvious solution would be for agents in such situations to promise to choose the cooperative option rather than the selfish one. But, for reasons already elaborated (in subsection 5.2), none of them is able to make such promises (not sincerely, at least); as a result, no cooperation occurs. If, however, the agents were agents of some other kind—constrained maximizers, for example—then they could credibly promise to cooperate, and they would end up with an optimal outcome rather than the sub-optimal equilibrium outcome to which egoists are condemned. So the universal acceptance of egoism can be expected to have worse consequences, even in terms of the agent's own interests, than the universal acceptance of some other theory of rationality. In such cases, then, egoism exhibits collective self-defeat.

7.3 Having considered promises, we turn now to threats. We saw above (in subsections 5.4, 5.6, and 5.8) that an egoist cannot exploit others, or defend himself against others, as well as a threat-fulfilling agent could, and that he could do better still if he were also a threat-ignorer. These results indicate that egoism is individually self-defeating in threat cases. Do similar results indicate that egoism is collectively self-defeating in threat cases? That is, would a group of egoists all benefit if they all become threat fulfillers? And benefit still more if they became threat ignorers?

The answers to these questions expose some interesting disanalogies between the individual acceptance of certain decision procedures and the collective acceptance of them. First, consider whether a group of egoists would all benefit if they all became threat fulfillers. Intuitively, we are likely to think that, however, advantageous it would be from the individual perspective for one to become a threat fulfiller, it could be very unpleasant to live in a whole society of threat fulfillers (even if, as implied, one were a threat fulfiller oneself). But are there any circumstances in which everyone might stand to gain from everyone's being a threat fulfiller?

Here is one scenario. Suppose that you now have some indivisible and perishable resource that I want, and later I will have some indivisible and perishable resource that you will want, and each of us would rather have the other's resource than his own. Now a trade would be a natural solution to this problem, but suppose that you have no way to be sure that I will give you my resource when it becomes available. Since I will have no incentive to keep my end of the bargain, you may doubt that I will do so, and you may refuse to give me your resource first. As egoists, we would have to settle for the sub-optimal outcome in which no exchange takes place. But suppose, instead, that each of us is a threat fulfiller, and suppose that circumstances are such that each of can make a threat that will cause the other to give up his resource, but cannot make a threat that will cause the other to return a resource once it has been seized. Then I will use a threat to get your resource from you, and later you will use a threat to get me resource from me. In this way, the making of (and yielding to) threats may function as an efficient reallocation procedure. That is, in the absence of an adequate market, threats may do the job. By being threat fulfillers, we will end up with a Pareto-optimal outcome, doing better than if we were egoists. So although such a case could, as we noted earlier, be dealt with by way of a *promise* to trade, we have just seen that it could also be dealt with by way of the disposition to fulfill threats.

So we can imagine situations in which everyone would benefit if they were all threat fulfillers instead of simple egoists. And if such situations were to occur frequently, then a disposition to fulfill threats might enhance collective pragmatic effectiveness and not just individual pragmatic effectiveness. But what if such situations do not occur very frequently? What if resources are allocated efficiently, in the sense that each of us benefits more from what he has than he would benefit from what the other has? Then the outcome will be worse if we make threats that cause a trade. So will we refrain from making such threats? Unfortunately not. For regardless of whether I refrain from using a threat to seize your resource or not, you will still have an incentive to use a threat to seize my recourse. Thus, I lose nothing by using a threat to seize your resource first, and we end up trading, as in the previous case. So our being threat fulfillers could cause us to end up with a sub-optimal outcome despite an efficient initial allocation of resources.

The foregoing considerations suggest that the desirability of egoists' becoming threat fulfillers will depend on the frequency with which cases arise in which their all having the disposition to fulfill threats will be necessary and sufficient for enabling them to obtain an Pareto-optimal consumption pattern from an inefficient initial allocation. And it is hard to see how such cases could arise frequently enough to offset the obvious disadvantages to everyone of their all being threat fulfillers. I conclude, then, that although the egoist's inability to fulfill threats is individually pragmatically disadvantageous, it seems unlikely to be collectively so.

7.4 What about the egoist's inability to *ignore* threats? We saw earlier (in subsection 5.8) that this is a source of individual self-defeat for egoism. Is it also a source of *collective* self-defeat, or does the disanalogy between individual and collective self-defeat exposed in the case of fulfilling threats extend to the case of ignoring them?

Clearly if everyone is an egoist, then the representative individual does not stand to gain from everyone's also being a threat ignorer, since if everyone is an egoist, then no one makes threats and the disposition to ignore them would be otiose. But what if everyone is a threat fulfiller? Then does the representative individual stand to gain from everyone's also being a threat ignorer? That depends, as in the case of threat fulfillment, on the nature of the situations that arise. The greater the prevalence of situations of the kind specified earlier—in which the disposition to fulfill threats is necessary and sufficient for the obtaining of a Pareto-optimal consumption pattern from an inefficient initial allocation—the less useful would be a universal disposition to ignore threats. Assuming (as above) that such situations can be expected to arise fairly infrequently, it seems likely that the egoist's inability to ignore threats will be not only individually self-defeating (as shown in subsection 5.8), but also collectively so.

This conclusion can be established in another, somewhat more elegant, way. Parfit points out that if everyone is a threat ignorer, then the advantages of being a threat fulfiller evaporate: "Since everyone else is now disposed to ignore my threats, they have become useless" (1984, p. 460). So if everyone is a threat ignorer, then the situation is essentially identical to one in which no one is either a threat fulfiller or a threat ignorer. Thus, the collective pragmatic effectiveness of a disposition to ignore threats is simply the inverse of the collective pragmatic effectiveness of a disposition to fulfill threats. It is no wonder, then, that the disanalogy between individual and collective self-defeat in the case of fulfilling threats does not extend to the case of ignoring them.

7.5 This section, as I said at the beginning of it, is devoted to exploring collective self-defeat as it arises from strategically induced dynamic inconsistency. Up to this point, we have explored this issue with respect to egoism. We now turn our attention to act consequentialism, which (as we will see) has been said to exhibit self-defeat of this sort. We shall find, however, that this specific criticism of act consequentialism is overstated at best.

There is a certain familiar sort of objection to act consequentialism that appears to show the collective self-defeat of that theory. This objection is that act consequentialists cannot be counted on to keep promises they've made and that, as a result, the practice of promising could not be sustained in a society in which act consequentialism is universally accepted. Brandt makes the point in the following way:

> suppose I have performed some service – mowed your lawn in response to a promise of payment of $10 for so doing. But when the service is completed, will I collect? The theory tells a person to do whatever will maximize utility by her expenditure. It seems that whether I am paid will depend on, say, your conception of the relative importance of various other purposes for which you might spend the money. When this occurs, the incentive to perform such services will be reduced, as will the incentive to make any plans that count on the predictable behavior of others. (1996, p. 144)

So it might appear that in a society of act-consequentialist agents, the practice of promising would be unsustainable. Moreover, the practice of truth-telling in general might appear to be unsustainable in a society of act-consequentialist agents. For such agents choose acts—including the saying of things—purely on account of their optimality and, as such, give no independent weight to being truthful in promising or in any other circumstances.

Underlying this general line of argument is a thicket of subtle and intricate issues that I shall not try to sort out here.[39] For I propose to estimate the strategically induced consequences (in terms of pragmatic effectiveness) of the universal acceptance of act consequentialism not by focusing on the sorts of interaction that would be *absent* in such a society (which is the orientation of the foregoing approach), but by focusing on the sorts of interaction that would be *present* in such a society.

[39] For discussions of them, see Hodgson, as well as Gibbard, Sobel, Barnes, Singer (1972), David Lewis, Mackie (1973), and Narveson (1976). For proposals of moral theories that act consequentialists should like, because of their improved pragmatic effectiveness in cases such as these, see Gauthier (1975b) and Regan.

There is a crucial feature of a society in which act consequentialism is universally accepted that distinguishes this case from every other that we have considered up to this point. In this case alone, all of the agents share a common aim. Although we do not know exactly what that aim is (because we have not said exactly what sorts of consequences the act consequentialism in question posits as *good* consequences), we do know that all of the agents aim at the same outcome, or at the realization of the same values in outcomes. We may regard them, then, as working as a *team* to promote certain values. And once we regard them in this way, we should not be surprised to find that there is no use for the practice of promising.

To see this, consider a team of workers assembling a car. Each worker does some part of the overall task; the workers communicate among themselves as to what they are doing and what they will do, and the car gets built. Promising has no place: in fact, it would be neither possible nor necessary for one worker to persuade another worker to perform a certain act by offering, as the incentive, to behave in a way that goes against his own preferences. This would not be *possible* because all of the workers have a common aim, meaning that the preferences of the first worker are shared by the worker whom he is trying to persuade. And she can hardly be persuaded to do something by the thought that if she does it, then he'll act against his own—and, by implication, her—preferences. And such persuasion would not be *necessary* because no worker would ever need to be persuaded by another worker to perform a certain act. For if it's in the interest of one worker for another worker to perform a certain act, then it's in the interest of that other worker, too, due to the exact coincidence of their interests. So it would appear that a society of act-consequentialist agents would have neither any possibly of, nor any use for, a practice of promising.[40]

[40] Does parallel reasoning show that *threats* would have no place in an act-consequentialist society? One disanalogy between promises and threats is that when a promise (of the sort under consideration) is kept, one party benefits at the expense of the other; but when a threat is followed through on, both parties lose. This commonality of interests already inherent in threat cases means that threats would be *possible* in an act-consequentialist society in a way that promises would not: specifically, if one agent somehow convincingly threatened another to inflict some harm on both of them unless she performs some act, then the coincidence of her interests with his would ensure that she would have an incentive to comply. (In contrast, recall that when one agent offers, as in a promise, to incur a harm in order to confer a benefit on another, the commonality of the agents' interests prevents such an offer from being effective.)

But although threats would be *possible* in an act-consequentialist society in a way that promises would not, they would still not be *necessary*. To see, this, recall the flood-control example and note that if the policymaker has an incentive to dissuade developers from building houses somewhere (such as in a flood plain), then the commonality of the agents' interests ensures that those developers have the same incentive not to build houses there. No threat, or any inducement at all, would be needed.

What about truth-telling? Unlike promising, truth-telling would be both useful and possible in an act-consequentialist society. To be sure, act-consequentialists agents would not regard lying as *inherently* wrong, but it is hard to see what incentives agents might have for cultivating false beliefs in one another that would outweigh the obvious incentives of enabling one another to have the information they need in order to advance the aims they share. And with the intermediate aim of cultivating true beliefs in view, act-consequentialist agents could hardly be blind to the superiority of truth-telling to any other way of cultivating true beliefs.

So a society of act consequentialists could accommodate truth-telling, and would operate just fine without a practice of promising. For such a society would be one in which all of the agents worked together, as a team, in order to advance a common aim or set of aims. It would appear, then, that act consequentialism does not suffer from problems of collective self-defeat, arising from strategically induced dynamic inconsistency, in the ways that many of its critics have thought.

7.6 The foregoing considerations suggest that the question of act consequentialism's *collective* self-defeat, as it arises from strategically induced dynamic inconsistency, must be answered quite differently from parallel question about egoism, and quite differently from the questions about the *individual* self-defeat of egoism and act consequentialism. For in the latter three cases (which we considered earlier, in section 5 and in subsections 6.2 through 6.4), strategic considerations—considerations pertaining to how agents' intentions and expectations interact—gave rise to problems of dynamic inconsistency that undermined pragmatic effectiveness. But in the case just considered (in subsection 7.5), no such complications arose. So whereas section 5 ended with the thought that both egoism and act consequentialism have obvious problems with individual pragmatic effectiveness due to strategically induced dynamic inconsistency, this section must end by noting that something analogous can be said in the collective case only about egoism, not about act consequentialism.

8 Conclusion

This chapter has been devoted to a consideration of the self-defeat (understood as pragmatic ineffectiveness) of straightforwardly maximizing theories. We began by considering problems of implementation (section 2), moved on to a brief look at problems that arise because of the sorts of desires that humans naturally have (section 3), and then dwelt at some length on self-defeat as it may arise from non-strategically induced dynamic inconsistency (section 4) and strategically induced dynamic inconsistency in the case of both individual (sections 5 and 6) and collective (section 7) acceptance of the theory being considered. And almost without exception (such as in subsection 7.5) our inquiries have served to identify several sources of self-defeat for straightforwardly maximizing theories. Although no attempt has been made to quantify the extent to which such theories may be pragmatically inferior to certain of their rivals, it seems fair to conclude that the pragmatic ineffectiveness of such theories is, even by the most conservative estimates, of sufficient magnitude to motivate an investigation into the significance of self-defeat. We undertake such an investigation in the next chapter.

III

There’s No Defeat in Self-Defeat

> though the philosophical truth of any proposition by no means depends on its tendency to promote the interests of society; yet a man has but a bad grace, who delivers a theory, however true, which, he must confess, leads to a practice dangerous and pernicious.
>
> —David Hume (1777), p. 279

1 Introduction

In the last chapter, we explored several ways in which straightforwardly maximizing theories, such as egoism and act consequentialism, can be self-defeating. We saw that it can be expected that, when agents subscribe to such theories—that is, when agents employ such theories as their decision procedures—the outcomes will typically be decidedly inferior (even in terms of the aims of those very theories) to the outcomes that would be brought about if the agents had employed certain other normative theories as their decision procedures. This chapter is devoted to assessing the significance of these results. I shall be seeking to draw two conclusions. The first is that the fact that a theory is self-defeating is not a good reason for rejecting it. The second is that, regardless of how significant self-defeat of this kind might be, many rivals of straightforwardly maximizing theories may also be criticized as self-defeating in the same way.

In order to establish these conclusions, I shall first, in section 2, present as strong an argument as I think is available for showing the significance of self-defeat. Then, in the section 3, I shall present and dispose of a familiar reply to the charge of self-defeat. In sections 4 and 5, I shall offer two more-successful replies to the charge of self-defeat; and in section 6, I shall show how self-defeat of the relevant kind besets not only straightforwardly maximizing theories, but also many of their rivals.

2 The case for the significance of self-defeat

2.1 Let us begin by exploring some very rudimentary intuitions for and against the significance of self-defeat. On the one hand, the thought that a theory ought not to be self-defeating enjoys a certain obvious plausibility. Indeed so intuitive do many authors seem to find this thought—or expect their readers to find it—that it is often stated with little or no argument. Hodgson, for example, writes with hardly any argument that

> the act-utilitarian cannot . . . readily admit that acting upon the act-utilitarian principle would have worse consequences than would acting upon the moral rules accepted in one's society, for this would make the principle self-defeating. (p. 3; see also p. 60)

And Parfit, though dismissive of the importance of self-defeat of several kinds, concedes that "If there is any assumption on which it is clearest that a moral theory should not be self-defeating, it is the assumption that it is universally successfully followed" (1984, p. 103).[1]

But why should the self-defeat of a normative theory be an ingredient in our evaluation of it? It is a philosophical commonplace that the good or bad consequences of an agent's holding a certain belief are a separate matter from the truth of the belief. For example, the fact that structural engineers may do their work more efficiently if they take a classical (i.e., Newtonian) rather than a relativistic (i.e., Einsteinian) approach to physics does not count as a reason to say that the Newtonian theory is the better account of physical phenomena; and the fact that better results may come from a defense attorney's believing that her client is innocent (and thereby being more motivated to provide him with the zealous defense to which he is entitled) does not count as evidence that her client really is innocent. Far, then, from agreeing with Mill's claim that "no belief which is contrary to truth can really be useful" (1859, p. 234 [ch II, par. 10]),[2] we acknowledge all the time that it can be useful for an agent to have false beliefs. In this same vein, instead of thinking that the self-defeat of a normative theory is a factor to be considered when evaluating it, we may think that situations sometimes arise in which irrational agents happen to fare better than rational ones, and that the situations discussed in the last chapter are of this kind—meaning that the soundness of a theory of rationality or of a theory of morality is not

[1] He adds, a few pages later, that "a moral theory must be collectively successful" (1984, p. 111).

[2] For comments on "the exaggerated emphasis Mill places on truth," see Crisp (p. 192).

impugned by its self-defeat.[3] Why, then, should self-defeat matter in the case of normative theories?

Here is one consideration that may help to explain, if not to justify, the force that self-defeat claims are often felt to have in the evaluation of normative theories: normative theories are not amenable to empirical testing in the (relatively) simple way in which, say, scientific theories and beliefs about certain past events (such as the role of a particular defendant in the commission of a particular crime) are. We have procedures for testing beliefs of the latter kind against evidence, and these procedures do not seem to work as well, if they work at all, in the evaluation of normative theories. To be sure, it is not *universally* agreed that normative theories cannot be held to answer in any way to the tribunal of ordinary observable facts—Harman, for example, tells a complex story about how moral facts may be on more solid epistemological footing than they may initially appear to be (pp. 130–32)—but at the same time there is no method for testing normative theories against observable facts that commands even *general* assent. In the absence of the usual dominant criterion for evaluating theories (fidelity to observable facts), other criteria—such as pragmatic effectiveness—are enabled to seem more appealing.

So the difficulty of testing normative theories in the way that we test scientific theories may open the door to the influence of self-defeat considerations in our evaluation of normative theories. But can we justify letting such considerations in? Are such considerations reasonable bases for discriminating among normative theories? A preliminary argument *against* allowing such considerations to have any influence grows out of the commonplace, expressed above, that the good or bad consequences of an agent's holding a certain belief are a separate matter from the truth of the belief. Developing this thought, it has been claimed that

> We can distinguish between two kinds of cognitive reasons, epistemic and pragmatic, based on different grounds. The ground of the former is that their constituent guarantees or makes it probable that its target belief is true, that of the latter, that having its target belief has certain desired consequences, independently of whether it is true or false. (Baier 1995, p. 80)

This distinction is, of course, familiar. When, in Hume's *Dialogues Concerning Natural Religion* (1779), Cleanthes argues that we ought to believe in the existence of God because that is the best

[3] As Scheffler puts the point, the self-defeat of utilitarianism "does not show that the utilitarian principle is irrational, any more than the principle that students ought to stay calm during important examinations would be proven irrational if it should turn out that students who made calmness their conscious aim ended up tenser than they would otherwise have been" (p. 46).

explanation of the order, complexity, and other apparent marks of design that we observe in the world around us, he is giving an epistemic reason; when he claims that "The doctrine of a future state is so strong and necessary a security to morals that we ought never to abandon or neglect it" (p. 82 [ch. XII, par. 10]), then he is giving a pragmatic reason. With this distinction between epistemic and pragmatic reasons explicitly before us, we should be a little less moved by the observation, mentioned earlier, that a good reason for allowing considerations of self-defeat to influence our evaluation of normative theories is that they cannot be tested against observable facts in the usual way. For when we allow considerations of self-defeat to have any influence, it may appear that we are not just replacing our standard epistemic reasons for beliefs with second-best or quirky epistemic reasons. It may seem that we are, instead, replacing epistemic reasons with pragmatic reasons.

To be sure, there are some purposes for which an appeal to pragmatic reasons would be justified. If, for example, we were interested in identifying the beliefs the having of which by us and by others would result in certain consequences that we might regard as desirable, then we would want to know about pragmatic reasons for and against certain beliefs. But this is not an inquiry of that sort. Like most standard inquiries in moral philosophy,[4] ours is an epistemic one, not a pragmatic one. So the argument cannot *just* be that straightforwardly maximizing theories are unacceptable because they tend to be self-defeating.

2.2 But the argument isn't just that—or, at least, it needn't be. Indeed a considerably more sophisticated case for the significance of self-defeat has been offered by Gauthier. Although Gauthier's arguments concern theories of instrumental rationality in particular, they are easily applicable to consequentialist theories of morality as well. According to Gauthier, the reason why the results—in terms of the effects on the furtherance of an agent's ends—of an agent's subscribing to a certain theory of rationality matter to the evaluation of that theory is simply that the furtherance of the agent's ends is what instrumental rationality is all about:

[4] Sidgwick, for example, writes that "here as in other opinions we ought to aim at nothing but truth" (p. 335) and "on reflection it is generally admitted that it cannot be good to be in error on this or any other point" (p. 429). And G. E. Moore is (unsurprisingly) more emphatic in his claims that "The direct object of ethics is knowledge and not practice" (p. 71) and that "What I am concerned with is knowledge only—that we should think correctly and so far arrive at some truth, however unimportant: I do not say that such knowledge will make us more useful members of society" (p. 115). A notable exception is Aristotle, who writes that "Our present inquiry does not aim, as our others do, at study; for the purpose of our examination is not to know what virtue is, but to become good" (p. 1103b, ll. 27–29).

> If a person's reasons take their character from her aim, then it is surprising and troubling if acting successfully in accordance with her reasons, she must sometimes expect to do less well in relation to her aim than she might. . . . If the orthodox account of the connection between aim and reasons were correct [in other words, if egoism were the best conception of rationality], then sometimes I should not expect success in acting on my reasons to lead to my life going as well as possible. And so I propose to rethink the connection. (1994, p. 694)

The connection that Gauthier proposes is the following:

> I conclude that deliberative procedures are rational if and only if the effect of employing them is maximally conducive to one's life going as well as possible. (1994, p. 701)[5]

To compare two rival theories of rationality, he explains, we are to consider the situations in which they would, as decision procedures, yield different choices, and if one set "would sometimes access inferior prospects [relative to the other], and never access superior prospects, then reject it as less than fully rational" (1997b, p. 22). It follows that we ought to reject egoism as less than fully rational, since we saw in the last chapter that the effects (for the agent) of being an egoist are sometimes worse than, and never better than, the effects of being a constrained maximizer or a resolute chooser. That is, we saw that constrained maximization and resolute choice each prescribe deliberative procedures that are more conducive to the furtherance of the agent's ends, or to her life's going well, than are the deliberative procedures prescribed by egoism. And according to the test for deliberative rationality just quoted, this matters: the pragmatic effectiveness of constrained maximization and resolute choice makes them better conceptions of rationality than egoism.

On this view, the very concept of rationality in deliberation and action is a complicated arrangement in which (1) rational deliberation is deliberation that follows the decision procedure (be it understood in terms of dispositions or rules or other decision sources) that is optimal for the agent in question and (2) rational action is action that is prescribed by rational deliberation. That is, rationality in deliberation and action is essentially a two-tier affair, with the rationality of decision procedures being at one level; with the rationality of acts being at another, subordinate, level; and with rationality at the first level "trickling down" to the second level in such a way that the rationality of decision sources may be transmitted to, or inherited by, acts.

[5] In a paper written earlier (but published later), he makes essentially the same point in this way: "a rational action is one motivated by a rational disposition, and a disposition is rational if and only if all of the actions it motivates collectively lead to the agent's life going at least as well as it would if she were motivated by any other possible disposition" (1997a, p. 34).

Before proceeding, we should pause to make clear the roles of the concepts of self-defeat and pragmatic effectiveness. The fact that egoism is self-defeating, in some sense stricter than that of simple pragmatic ineffectiveness, is actually something of a red herring. What matters is that egoism, as a decision procedure, is not pragmatically effective, in that egoists fare worse than constrained maximizers and resolute choosers. The fact that egoism sets pragmatic effectiveness as the aim of conduct (making egoism self-defeating in some sense stricter than that of simple pragmatic ineffectiveness) is, at the end of the day, a coincidence. After all, a theory of rationality can easily be pragmatically ineffective without being self-defeating in this stricter sense. For example, an agent who regards it as rational always to choose whatever conduct will minimize her exposure to the sun may actually succeed in minimizing her exposure to the sun, meaning that her theory of rationality is not self-defeating in this stricter sense; but unless she has a very unusual set of ends, with avoiding exposure to the sun being pre-eminent among them, then on the whole the effect of employing such a decision procedure is likely to be worse than the effect of employing a decision procedure that is somewhat more responsive to the ends that she actually has. Such a theory of rationality would not be self-defeating in this stricter sense, but it would clearly be pragmatically ineffective, and it is on *this* basis that the view being discussed here would reject it. Similarly, the problem with egoism is not that it is self-defeating in this stricter sense (though it does seem to be),[6] but that it is pragmatically ineffective. For our purposes, it is just a coincidence that the pragmatic ineffectiveness of egoism also makes it self-defeating in this stricter sense. This is why I said above (in section 1 of chapter II) that self-defeat was to be understood simply in terms of pragmatic ineffectiveness.[7]

Understanding self-defeat as pragmatic ineffectiveness enables our discussion of the significance of self-defeat to bear on McClennen's defense of resolute choice. For McClennen is unequivocal in insistence on the significance of pragmatic effectiveness. McClennen writes, for example, that "Any argument that succeeded in grounding [principles of choice] in . . . explicitly pragmatic considerations would come as close as one might hope to being dispositive" (1990,

[6] Scheffler, though, questions the propriety of using the term 'self-defeat' to characterize the phenomenon under discussion (p. 45).

[7] Thus what I refer to as self-defeat does not include a certain phenomenon that Kavka regards as a form of self-defeat: that of a theory's being self-effacing, or requiring its own abandonment (1986, p. 365, n. 64). Although we shall consider such phenomena at length in the next chapter, we shall not refer to them as instances of self-defeat.

pp. 10–11). And what McClennen means by "pragmatic considerations" parallels (indeed, anticipates) the remarks by Gauthier quoted above:

> the thrust of the "pragmatic" argument is that one who fails to satisfy certain choice conditions will end up with a less preferred consequence than he could have achieved if his choice behavior always satisfied the conditions in question. (1990, p. 84)[8]

He then defends resolute choice precisely in these terms, referring to "The pragmatic impeccability of resolute choice" (1990, p. 198).[9]

2.3 So Gauthier's argument for constrained maximization, along with McClennen's argument for resolute choice, rest on the pragmatic effectiveness of their theories, relative to egoism. But exactly how is pragmatic effectiveness supposed to matter? It is often thought that Gauthier's defense is a sort of counterfactual pragmatic one: that it is rational to deliberate and to act in the manner of a constrained maximizer simply because, *if* they had the chance, then egoists *would* choose to become constrained maximizers.

To be sure, there are passages that invite this reading of Gauthier. For example, in his 1975 paper "Reason and Maximization" (1975a), he defends constrained maximization in just this way (pp. 228–230). Similarly, in the chapter defending constrained maximization in his 1986 book *Morals by Agreement*, he writes,

> To demonstrate the rationality of suitably constrained maximization, we solve a problem of rational choice. We consider what a rational individual would choose, given the alternatives of adopting [egoism], and of adopting constrained maximization, as his disposition for strategic behaviour. . . . Thus he compares the expected utility of disposing himself to maximize utility [i.e., the expected utility of adopting and then continuing to accept egoism] . . . with the utility of disposing himself to co-operate with others [in the manner of the constrained maximizer]. (pp. 170–171)

And some authors seem to take remarks such as these as indicative of the essential character of Gauthier's argument. Velleman, for example, writes that "The only authority available to a principle of practical reasoning [such as constrained maximization], in Gauthier's picture, lies in the fact that the conception is supported by itself or by another conception [such as egoism]" (p. 230).

[8] See also DeHelian and McClennen, in which a certain approach is said to "fai[l] the most basic of all pragmatic tests" because "those who reason in the way [it] recommends do less well than those who reason differently" (p. 329).

[9] See also his 1997, p. 234, p. 246, and p. 250, as well as his 1988, pp. 112–113. See also Broome, who writes, "That resolute choice can be advantageous is McClennen's main reason for thinking it rational" (p. 668). ("But it is an insufficient reason," Broome adds [p. 668].)

Now the foregoing argument for constrained maximization is easily refuted. As Parfit argues (1984, pp. 19–20), either egoism is a theory whose prescriptions ought to be regarded as worth following, or it isn't. If it is, then this fact alone should enable it to be regarded as the best theory of rationality; and if it isn't, then the fact that it directs agents to subscribe to some other theory of rationality can hardly be regarded as a source of *support* for that other theory. Now it has been claimed—by, e.g., Richard Dean (p. 465, n. 11)—that matters are more complicated than this simple dilemma indicates, because the choice of what theory to adopt is a parametric one, not a strategic one, and constrained maximization agrees with egoism when it comes to parametric choice. Thus, constrained maximization is recommended not only by egoism, but also by itself. But Parfit's objection can be modified to take the form of Velleman's complaint: "Why should we feel bound by principles whose only claim on us is that they are recommended by themselves or by other principles that have even less to recommend them?" (p. 230).

2.4 But the failure of the foregoing argument does not tell against Gauthier's argument, because it rests on an interpretation of Gauthier's approach that takes too seriously his talk of *choosing* a theory of rationality (as signaled, perhaps, by the title of Velleman's paper: 'Deciding to Decide'). As Gauthier writes,

> We have defended the rationality of constrained maximization as a disposition to choose by showing that it would be rationally chosen. . . . But the idea of a choice among dispositions to choose is a heuristic device to express the underlying requirement, that a rational disposition to choose be utility-maximizing. We may therefore employ the device of a parametric choice among dispositions to choose to show that in strategic contexts, the disposition to make constrained choices, rather than straightforwardly maximizing choices, is utility-maximizing. We must however emphasize that it is not the choice itself, but the maximizing character of the disposition in virtue of which it is choiceworthy, that is the key to our argument. (1986, p. 183)

So, the fact that a disposition is choiceworthy from the point of view of egoism may be *evidence* that that disposition is rational, but it would not be the *source* of the rationality of that other disposition. In other words, the fact that a disposition is choiceworthy from the point of view of egoism is only indicative, not constitutive, of its being rational. As Dean writes, "Nowhere does Gauthier accept [egoism] as the correct rational policy, then argue that this policy recommends choosing to follow some other policy" (p. 459). How can we state Gauthier's argument so that the key to the argument is not the choiceworthiness of constrained maximization, but "the maximizing character of [it] in virtue of which it is choiceworthy"? Drawing on Gauthier's later remark (quoted earlier) that "deliberative procedures are rational if and only if the effect of employing them is maximally conducive to one's life going as well as possible" (1994, p. 701), we can state his argument in the following way:

(P1) The best theory of rationality—the theory such that deliberation and action in accordance with it are rational—is the most pragmatically effective one.

(P2) The most pragmatically effective theory of rationality is constrained maximization.

(C) Therefore, the best theory of rationality is constrained maximization.

Thus stated, it is clear that the justification for constrained maximization need not refer to an egoistic choice of dispositions. As Dean writes, "Constrained maximization is the correct policy because it is rational (i.e., it provides the most utility), not because it is rationally *chosen*" (p. 459).

The argument is obviously valid, and we shall assume for the sake of discussion that the second premise is established by the results of the last chapter. But what can be said in support of the first premise? This statement, which I call the *pragmatic-effectiveness criterion*, is the most plausible basis for assigning any significance to results such as those obtained in the last chapter, so it is crucial now to assess its merits.

Gauthier defends the pragmatic-effectiveness criterion by appealing to intuitions such as this one: "it is surely mistaken to treat rational deliberation as self-defeating, if a non-self-defeating account is available" (1994, p. 702). But in noticing Gauthier's reliance on such intuitions it is easy to make the mistake of thinking that Gauthier's argument for the pragmatic-effectiveness criterion is a pragmatic one. That is, it is easy to think that Gauthier thinks that this criterion is an appropriate one for evaluating theories of rationality because an agent who evaluates a theory of rationality by employing this criterion can expect to do better than one who does not. (For an agent who accepts this criterion for evaluating theories of rationality will be led to adopt and then to continue to accept constrained maximization or resolute choice or some theory of rationality like that, and such an agent will do better than one who denies this criterion in favor of the orthodox outlook, which leads to egoism.)

Velleman, for example, quotes Gauthier's remark about conceiving of rational deliberation as non-self-defeating, if possible, and then claims that "Gauthier has argued in a circle" because the only reason he gives for his "pictur[e] of how conceptions of practical reasoning should be evaluated"—the only reason Gauthier gives for the pragmatic-effectiveness criterion—is that the orthodox approach, by supporting egoism, "recommends a self-defeating" theory of rationality (p. 228). Velleman goes on to say that "We have to join Gauthier in adopting this picture, then, before we can see it as saving us from a self-defeating" theory of rationality (p. 229). If this last remark of Velleman's means that have to join Gauthier in taking a pragmatic perspective before we can see the merits of the pragmatic-effectiveness criterion, then it implies that Gauthier not only defends constrained maximization in terms of pragmatic effectiveness, but then defends that pragmatically-oriented defense of constrained maximization in terms of pragmatic effectiveness.[10]

The pragmatic-effectiveness criterion cannot, however, be adequately defended in this way. For an appeal to the pragmatic effectiveness of that criterion just pushes the issue back one level, and invites the question of why the most pragmatically effective criterion for evaluating theories of rationality should be regarded as the right one. What is still missing is an *epistemic* reason for regarding, as authoritative, either an indirectly maximizing theory of rationality (such as constrained maximization or resolute choice) or a criterion (such as, of course, the pragmatic-effectiveness criterion) that supports such a theory. The argument cannot be pragmatic all the way back.

[10] Velleman also writes that "Nothing underwrites our principles of practical reasoning in his picture, other than the principles themselves" (p. 233). Despite the textual evidence, however, my interpretation of Velleman must be regarded as tentative, for three reasons. First, it is not clear that Velleman entertains the possibility that Gauthier's argument is anything other than an appeal to egoism of the sort from which it is crucial to *distinguish* the pragmatic-effectiveness criterion. So I may be attributing, to Velleman, criticisms of an argument that he does not, in fact, consider. But Velleman's comments appear to be applicable to Gauthier's appeal to the pragmatic-effectiveness criterion, especially since the remark about the desirability of a non-self-defeating theory of rationality, which triggers Velleman's critical reaction, is from "Assure and Threaten," in which Gauthier is quite explicit in appealing not to an egoistic choice of dispositions, but to the pragmatic-effectiveness criterion.

The second problem with my interpretation on Velleman is that the passage of his from which I draw here is infused with a stricter, and more literal, notion of self-defeat—on which a theory cannot be self-defeating "unless its own evaluation is an inquiry of the sort to which the [theory] itself applies" (p. 228)—than I believe Gauthier uses, and than I specify earlier (where I say that the self-defeat of a normative theory is to be understood simply in terms of pragmatic ineffectiveness).

Third, if Gauthier's defense of the pragmatic-effectiveness criterion were itself pragmatic, then it would not exactly be circular, as Velleman says Gauthier's argument is: since the pragmatic-effectiveness criterion is a criterion for evaluating *theories of rationality*; it cannot also function—as it would need to function if it were to be defended by an appeal to *itself*, as would be needed for genuine circularity—as a criterion for evaluating *criteria for evaluating theories of rationality*. (That is, if Gauthier *were* to defend the pragmatic-effectiveness criterion in pragmatic terms, he would need to introduce and to appeal to another, higher-order, pragmatic-effectiveness criterion, such as one that said that the appropriate criterion for evaluating normative theories is the one the employment of which [in evaluating theories of rationality] would be most pragmatically effective.) But Velleman's charge of circularity can be made sense of as the claim that Gauthier's argument is circular insofar as an appeal to pragmatic considerations in the evaluation of theories of rationality (i.e., the pragmatic-effectiveness criterion) is itself justified in pragmatic terms.

2.5 But in Gauthier's remarks on constrained maximization and the significance of its pragmatic effectiveness, we can discern an argument that is not entirely pragmatic. For when Gauthier says what Velleman quotes him as saying (that "it is surely mistaken to treat rational deliberation as self-defeating, if a non-self-defeating account is available"), what he means is that we have a theoretical choice to make about what criterion for the evaluation of normative theories to regard as the correct one, or about what criterion for evaluating theories of rationality makes the most sense; and what Gauthier is claiming is that it militates in favor of a criterion for evaluating theories of rationality if that criterion does justice to our intuition that rational deliberation is not self-defeating. Of course, this does not mean that *any* theory of rational deliberation on which rational deliberation is non-self-defeating is superior to *any* theory of rational deliberation on which it is self-defeating (nothing in Gauthier's account commits him to the view that the non-self-defeating character of rational deliberation is an incontrovertible intuition); but only that it matters.

One consideration in favor of this intuition is that, according to the broadly consequentialist framework within which we have been proceeding, such an intuition may seem fairly benign: after all, what can be more rational for an agent than for him to make decisions on the basis of reasons which are such that his choosing on the basis of *them* furthers his ends as much as would his choosing on the basis of any other reasons? And, to extend the thought to theories of morality, what can be more morally acceptable than for agents to make decisions on the basis of reasons which are such that their choosing on the basis on *them* results in outcomes that are as good as those that would result from their choosing on the basis of any other reasons?

Gauthier's argument, then, should be construed in the following way. If we approach the evaluation of theories of rationality in the way suggested by the pragmatic-effectiveness criterion, then we can do justice to several of the intuitions that we have about rational deliberation. These intuitions may, of course, be at varying levels of concreteness and abstractness. One that is fairly intermediate between the concrete and the abstract has been referred to: that rational deliberation is not self-defeating. Another intuition, related to the first one but much more concrete, is that "if one enters an agreement rationally, then one must act rationally insofar as one acts in accordance with the agreement, at least if the circumstances remain as one envisages them in entering the agreement" (1975a, p. 225; see also 1984c, p. 159, and 1994, pp. 704–707). As we saw in the last chapter, this intuition must be rejected by straightforwardly maximizing theories; but it would appear from constrained maximization and resolute choice that theories selected by the pragmatic-effectiveness criterion can be expected to endorse it.

A very abstract intuition, and perhaps the one that lies at the root of Gauthier's defense of the pragmatic-effectiveness criterion, is signaled by Gauthier's reference, at the end of "Assure and Threaten," to a memorable passage from Nietzsche's *Genealogy of Morals*. Nietzsche claims that when we consider the development of humanity over time,

> we discover that the ripest fruit is the *sovereign individual*. . . . This emancipated individual . . . this sovereign man—how should he not be aware of his superiority over all those who lack the right to make promises and stand as their own guarantors, of how much trust, how much fear, how much reverence he arouses . . . and of how this mastery over himself also necessarily gives him mastery over circumstances, over nature, and over all more short-willed and unreliable creatures? (pp. 59–60)

And this intuition is, as we have seen, only one of several that may be adduced in support of the pragmatic-effectiveness criterion.

So Gauthier can be read as defending the pragmatic-effectiveness criterion not on pragmatic grounds, but on epistemic grounds. That is, his point is *not* that an agent who regards the pragmatic-effectiveness criterion an appropriate criterion for evaluating theories of rationality will do better than he otherwise would (though this may be true); his point is that a *believer*—such as a philosopher—who so regards the pragmatic-effectiveness criterion will then be able to conceive of rational deliberation and action in a way that embraces, rather than repudiates, several intuitively compelling thoughts. In other words, Gauthier maintains the pragmatic-effectiveness criterion is true regardless of whether it helps an agent to believe it. Underlying Gauthier's argument, it is crucial to see, is a conception of what a rational agent *is*, not just a thesis about what criterion for evaluating theories of rationality it would *profit* one to regard as authoritative. In short, there are certain intuitions that we have about rational agents, plans, and actions, and Gauthier's theory makes more sense of these than does the orthodox approach, with its embrace of egoism.

2.6 Or so it might be claimed. For in the next few sections, I shall offer a series of arguments, each more forceful than its predecessor, that cast doubt on the soundness of the pragmatic-effectiveness criterion. While some will focus on theories of rationality, others will focus on theories of morality, since it should be clear how the discussion undertaken so far can be extended to provide an argument for a pragmatic-effectiveness criterion for theories of morality on which the best theory of morality—the one such that deliberation and action in accordance with it are (morally) right—is the one such that its general acceptance in a society would advance that society's aims more than would the general acceptance of any other theory of morality. Throughout, I shall assume that the fortunes of instrumental theories of rationality and consequentialist theories of morality are sufficiently closely linked for claims for and against the pragmatic-effectiveness criterion for theories of rationality to imply, and to be implied by, parallel claims for and against the pragmatic-effectiveness criterion for theories of morality.

Before introducing any of those arguments, though, I would like to describe and set aside one strategy for attacking the significance of self-defeat that I shall *not* be pursuing. In the last chapter, I distinguished *individual* self-defeat—the sort of self-defeat that can arise when just one agent subscribes to a certain normative theory (whether a theory of rationality or a theory of morality)—from *collective* self-defeat—the sort of self-defeat that can arise when every agent in a group of agents subscribes to a certain normative theory (again, of either kind). One strategy for downplaying the problem of self-defeat is the piecemeal one of arguing that these two sorts of self-defeat—individual and collective—are not equally damaging to theories of rationality and theories of morality, respectively. It might be claimed, for example, that only individual (and so not collective) self-defeat could possibly be damaging to a theory of rationality; and that only collective (and so not individual) self-defeat could possibly be damaging to theories of morality. But although claims to this effect have been made[11] (and although I am sympathetic to such claims), I shall not seek to derive any advantage from them here. Instead, my strategy will be to attack the very idea that self-defeat of either kind is a reason to reject a normative theory of either kind. And as I said above, in what follows, certain arguments will be more conveniently expressed in terms of theories of rationality, while certain others will be set out in the context of theories of morality; but except where otherwise indicated, the following discussion is meant to apply both to individual and to collective self-defeat, and both to theories of rationality and to theories of morality.

[11] See, for example, Gauthier (1984a, p. 255) and Kavka (1986, p. 109, pp. 112–113, and pp. 271–272).

3 The reply from self-effacement

3.1 A familiar reply from defenders of egoism and act consequentialism to the charge of self-defeat is to claim that if being a straightforwardly maximizing agent (whether in rationality or morality) were self-defeating for some agent in some circumstance, then the agent would be enjoined, by the very straightforwardly maximizing theory in question, to become an agent of a different sort—an agent who employed a pragmatically optimal decision procedure. For example, an egoist might be led to become a constrained maximizer or a resolute chooser, and an act-consequentialist might be led to subscribe to some form of rule utilitarianism or common-sense morality.

Now it might be thought that such self-effacement is intrinsically objectionable and intolerable, so that even if it would save a theory from being criticized on grounds of pragmatic effectiveness, then it would also expose that same theory to dispositive criticism from another quarter. And I shall consider this and related claims in the next chapter. This section is devoted to an examination of whether self-effacement, however damaging it might be thought to be on other grounds, at least answers the charge of self-defeat.

3.2 One objection to the self-effacement reply highlights the well-known paradoxes that surround the idea of getting oneself or others to acquire and retain beliefs for pragmatic, rather than epistemic, reasons. Suppose that an egoist wants to become, say, a constrained maximizer. In order to do this, he must walk a fine line between two perils. On the one hand, his commitment to egoism must be strong enough to get him to regard his task as genuinely advisable: he must regard egoism as authoritative enough to give him genuine reasons for acting. On the other hand, his commitment to egoism must be weak enough for him to dislodge it somehow and to replace it with a commitment to constrained maximization. Furthermore, even assuming that the agent is constituted (with a commitment to egoism of the right intensity) so that he can walk this tightrope, a further problem will confront him: it is probably true that his commitment to constrained maximization would be shaken if he were ever to recall his reasons for becoming a constrained maximizer, since people tend not to regard pragmatic reasons as good reasons for holding certain beliefs (though they may, of course, regard pragmatic reasons as good reasons for trying to *cause* themselves to acquire and to retain certain beliefs). As Elster writes,

> Beliefs, like courage, can be instrumentally useful and yet be out of reach for instrumental rationality. It is part of the notion of a belief that to believe something implies the belief that one holds the belief for a reason, and not merely because of the utility of holding that belief. Beliefs are judged by their ancestors, not by their descendants. . . . it may often be the case that a necessary condition for achieving a little is the mistaken belief that one will accomplish much, but this could not serve as a *reason* for having excessive self-confidence. Bootstrap-pulling works only when one looks the other way. (1983, pp. 51–52)

As a result, the agent will face what Elster calls the "*problem of the self-eraser*" in that "The decision to believe [constrained maximization] . . . will hardly have any impact unless the person can bring himself to forget that his belief is the result of a decision to believe" (1983, pp. 57–58).

So, the mechanics of an agent's turning himself into an agent of another sort are far from trivial. And of course it cannot be supposed to be doable by the agent's flipping a switch, or by any single act of will. But more roundabout techniques may be viable. Kavka, referring to an agent who seeks to become a threat-fulfiller, writes that

> he must seek to initiate a causal process (e.g. a reeducation program) that he hopes will result in his beliefs, attitudes, and values changing in such a way that he can and will have the intention to apply the [retaliatory] sanction should the offense be committed. Initiating such a process involves taking a rather odd, though not uncommon attitude toward oneself: viewing oneself as an object to be molded in certain respects by outside influences rather than by inner choices. (1978, p. 296)

The distinction between "outside influences" and "inner choices" is crucial to the plausibility of self-transformation of the kind in question. In John Heil's words, although "the forming of beliefs, considered in itself, seems not to be under the immediate control of the will" (1983a, p. 359), nevertheless "it seems uncontroversial to allow that one may voluntarily stake steps that will eventually lead to the formation of particular beliefs" (1983b, p. 753).[12] Moreover, the possibility of self-transformation of this sort may be as important to the critics of straightforwardly maximizing theories as to their defenders, since the feasibility of revisionist theories depends on persons' having some control over this aspect of themselves. Gauthier, for

[12] See also Heil (1984 and 1992), along with Williams (1973), Winters, and Stocker (1982). In addition, see Cohen (pp. 282–83) and Bertram (pp. 27–33). Cohen likens self-transformation of this sort to the "remarkable change in man, by substituting justice for instinct in his behavior" to which Rousseau refers (quoted in Cohen, p. 282).

example, explicitly argues for the feasibility of an egoist's becoming a constrained maximizer (1975a, pp. 230–231).[13]

3.3 Assuming, then, that the self-effacement of a theory must be acknowledged to be feasible, what is the upshot for the charge of self-defeat? The thought is that once an agent of one kind is transformed into an agent of another kind, then the agent's original decision procedure can no longer be faulted for being less pragmatically effective than that other decision procedure. For example, if egoism would direct an egoistic agent to become an agent of another sort, then the pragmatic effectiveness of egoism can no longer be measured by looking at the pragmatic success enjoyed by an agent who is disposed to act and choose egoistically *on an ongoing basis*. For when the egoist becomes, say, a constrained maximizer, then the pragmatic effectiveness of egoism must be regarded as equal to that of constrained maximization. Of course, it may only be fair to charge some "transformation costs" to egoism's account, but this sort of disadvantage of being an egoist hardly seems to be what defenders of the pragmatic-effectiveness criterion have in mind when they criticize egoism. In any case, such costs can be expected to be negligible relative to the costs and benefits that an agent experiences over the course of a life. The upshot will be that, even on the metric employed by the pragmatic-effectiveness criterion, egoism will fare only negligibly worse than constrained maximization or resolute choice. As a result, the pragmatic-effectiveness criterion does not provide a reason for rejecting egoism or act consequentialism, since these theories contain (on any reasonable interpretation of them) provisions for self-effacement if circumstances warrant. To suppose otherwise, it might be said, is to evaluate an impoverished or badly implemented form of egoism: one that either does not allow for its own effacement or that is implemented by agents who do not notice the desirability of transforming themselves even when egoism would so instruct them.

3.4 But the pragmatic-ineffectiveness criticism of egoism does not presuppose either an impoverished form of egoism or that it is badly implemented. On the contrary, we may suppose that egoistic agents successfully turn themselves into whatever type of agent it would be best for them to turn themselves into. Indeed we may suppose that such self-transformation is such that the long-term prospects of an agent who is now an egoist (long-term prospects in terms of his life's going well) are only negligibly worse than those of an agent who now subscribes to

[13] For further discussion of the feasibility of constrained maximization, see McClennen (1988, pp. 104–108) and Gauthier (1988, p. 212).

the most pragmatically effective theory of rationality. Still, this does not enable egoism to deflect the pragmatic-ineffectiveness criticism. For the charge is not that someone who is *now* an egoist would do worse, in the long run, than someone who is now a constrained maximizer or a resolute chooser. The claim is that someone who lives his whole life as an egoist would do worse than someone who lives her whole life as a constrained maximizer or as a resolute chooser.

This brings out an important feature of the idea of assessing a normative theory's pragmatic effectiveness. Running the pragmatic-effectiveness test on a normative theory involves holding *fixed* the corresponding decision procedure over the course of a life. That is, we suppose that an agent faithfully follows his decision procedure *except* to the extent that his doing so would result in his becoming an agent who ceases to follow that decision procedure. So, we allow the egoist to do all sorts of maximizing things; we even allow him to engage in precommitment activities and to take other steps to avoid or to compensate for the unfortunate effects of his being an egoist; but we do not (for the purposes of ascertaining egoism's pragmatic effectiveness) allow him to cease to be an egoist.

Now it might be objected that if a theory contains self-effacement provisions (as egoism and act consequentialism surely do), then these should be allowed to have effect. To claim otherwise is to assume that a self-effacing egoism is no more pragmatically effective than a non-self-effacing egoism, which certainly *cannot* be assumed. It is simply misleading, and unfair, (it might be objected) to hold fixed the agent's decision procedure just to preserve the conditions of the experiment.

But, to repeat, the pragmatic-effectiveness criticism of egoism is not that the long-term prospects of an agent who is *now* an egoist (long-term prospects in terms of his life's going well) are only negligibly worse than those of an agent who now subscribes to the most pragmatically effective theory of rationality. Rather, the pragmatic-effectiveness criticism of egoism is that an agent who lives his *whole life* as an egoist would do worse than someone who lives her whole life as a constrained maximizer or as a resolute chooser.

To see the point of proceeding in this way, imagine that you buy a new car. It's a big, luxurious, boat of a car, with all sorts of features such as computerized maps, sensors that track the inner workings of the car, and monitors that track external factors such as the price of gas and the frequency of repairs. And let us say that, for whatever reason, you use the car for a lot of long and frequent commutes, resulting in lots of stop-and-go travel that leads to bad mileage and lots of repairs. Now suppose the car is intelligent enough to advise you that, since it is costing you so much money for gas and repairs, you would do better, in the long run, to get a tiny little gas-efficient car. And suppose this is what you do, saving lots of money over the next several years.

Now, to be sure, you would certainly appreciate a car that would be "honest" enough to tell you when it was unwise for you to hang onto it. (Would that all repair shops were.) But would you credit the first car with the pragmatic benefits that you began to enjoy only upon replacing it with the second car? To answer this question, consider the advice you would give a friend who has identical driving habits and who asks you which car he should get. Would you tell him that it doesn't matter, because they're equally effective? Surely that would be a strange answer. It would make much more sense to say that he should get the second car, though you could add that even if he gets the first car he won't end up too badly off because the first car tells drivers for whom it's inappropriate that they ought to get the second car. Similarly, if one normative theory enjoins an agent to adopt some other theory, that theory cannot take credit for the pragmatic effectiveness of that other theory. At most, it can take credit for advising the agent to adopt that other theory.

3.5 Here is another, somewhat abstract but happily brief, way of seeing the inadequacy of the self-effacement reply. According to the pragmatic-effectiveness criterion, the best theory of rationality is the one that identifies, as reasons, those considerations which are such that, if an agent regards those considerations as reasons, then the agent's life will go as well as possible. To be sure, a self-effacing egoism would tell an agent to regard, as reasons, those considerations which are such that, if he regards those considerations as reasons, then the agent's life will go as well as possible. But *it* does not *identify* those considerations as reasons: it only tells the agent to *regard* them as reasons, or to set in motion some sequence of events that will transform him into an agent who regards those considerations as reasons. The considerations that egoism identifies as reasons are considerations that it does not help an agent to regard as reasons. Because egoism does not identify the most pragmatically effective considerations as reasons—but only tells agents to get themselves to regard them as reasons—it is not the theory selected by the pragmatic-effectiveness criterion.

4 The reply from counterexamples

4.1 So the pragmatic-effectiveness criterion says that the best conception of rationality for an agent is the one that, if he subscribes to it, makes his life go best: in other words, rational acts are those prescribed by the decision procedure the having of which by the agent is optimal. Gauthier, as we saw, adumbrates this criterion in papers published in 1994 and 1997. But more recently, he has acknowledged the force of a certain objection to this criterion, and he has refined his theory to try to defend it against this objection. In this section, I shall develop this objection and the refinement Gauthier introduces in reply to it; I shall argue, moreover, that Gauthier's refinement achieves, at best, only a stalemate against the objection.

The pragmatic-effectiveness criterion identifies rational deliberation and action, for some agent, as deliberation and action that accord with whatever decision procedure is optimal for that agent. Now it can be inferred from the results of the last chapter that almost invariably, an agent's optimal decision procedure will yield prescriptions that occasionally conflict with the purpose that that decision procedure is supposed to serve, and generally does serve. (Otherwise, an agent's optimal decision procedure would never yield prescriptions that differ from those of a straightforwardly maximizing theory, which the last chapter suggests is unlikely at best.) A more ominous way of expressing this point is to say that there is *no particularly systematic connection* between the contents of some agent's optimal decision procedure and the purpose to which that decision procedure is subservient. An agent may be so situated—in terms of his own perceptions of the world around him, in terms of his own reasoning abilities, and in terms of other agents whose interests may not coincide with his—that his optimal decision procedure makes no direct reference to the purpose that it is supposed to serve. This is not to say that it is badly constituted; on the contrary, by hypothesis this decision procedure is the agent's optimal one. The point is that the optimal way for the agent to achieve his preferred outcomes may be to aim at outcomes that are entirely different from, or even incompatible with, his preferred outcomes. But if an agent were, in effect, *completely oblivious* in his deliberations to his preferred outcomes, the rationality of such deliberation would be seriously in doubt.

The objection, then, is that the pragmatic-effectiveness criterion may, in certain circumstances, identify (as rational or moral) certain deliberative procedures that cannot reasonably be regarded as rational or moral. To clarify this phenomenon, let me borrow an illustration of it from an unlikely ally: Gauthier, who relies heavily on the pragmatic-effectiveness criterion in his defense of his theory, but who introduces the following illustration in order to motivate a revision in his theory that will, he claims, make it less vulnerable to objections such as the one I am now developing. We are to consider a case in which deliberation that accords with my optimal decision procedure is not directed at the achievement of my aims: when I am thinking about the factors that my optimal decision procedure tells me to think about, I am not thinking about my aims at all. To imagine a case in which this might be so, Gauthier invites us to suppose that I am under the control of a being, whom I call the tyrant, who will punish me if I deliberate in the usual way—with a view to advancing my aims—but will reward me if I "take her directives as reasons for acting in themselves, independently of how they relate to my concerns" (1998, p. 49). In such a situation, it would clearly be most effective—maximally conducive to the furtherance of my ends—for me just to forget about my ends and simply take the tyrant's directives themselves as reasons for acting.

But could such deliberation reasonably be regard as rational? The realm of the tyrant is such a weird place that successful deliberative procedures would be so bizarre— taking the tyrant's directives in themselves as reasons for acting, independently of how they relate to independently held ends—that only the most dogmatic advocate of the pragmatic-effectiveness criterion could endorse them as rational. Cases such as that of the tyrant, then, underwrite my first objection to the claim that the rational agent is simply the agent who does what it takes to succeed—or, more precisely, who thinks in the way in which an agent has to think in order to succeed. The objection is simply that cases of this kind constitute counterexamples to the pragmatic-effectiveness criterion by showing that pragmatic effectiveness can reside in the most bizarre, and (more to the point) patently irrational, of deliberative procedures; and this fact shows that pragmatic effectiveness is not indicative, much less constitutive, of rationality. The most pragmatically effective decision procedure for an agent is *not* necessarily an acceptable theory of rationality.

4.2 What is Gauthier's reply to this objection? Instead of rebutting it, he embraces it as an occasion for a refinement to his view. Acknowledging that the rational agent is not *simply* the agent who deliberates in the way in which an agent must deliberate in order to succeed—claiming, that is, that mere effectiveness is not enough—Gauthier allows that in order to be rational, an agent's deliberation must also be *directed* at the furtherance of her ends. In the case of the tyrant, effective deliberation is clearly not so directed. Rather, it's directed at the fulfillment of the tyrant's directives, with no regard for the agent's own aims. Rational deliberation would be ineffective, while effective deliberation would be irrational.

But Gauthier does not, of course, entirely banish considerations of pragmatic effectiveness from his approach to evaluating theories of rationality. Rather, he restricts their scope by making them lexically subordinate to another condition. He hints at this lexically prior condition in the following passage, which appears just before his discussion of the tyrant case:

> To guard against misunderstanding my account of deliberation, it is essential to emphasize that deliberative reasons relate to effective direction. They are not simply whatever considerations would need to weigh with someone if he is to realize his concerns. (1998, p. 49)

And to this Gauthier adds the following note, which refers to a passage quoted above (in subsection 2.2):

> Thus, what I said in another essay – 'deliberative procedures are rational if and only if the employing of them is maximally conducive to one's life going as well as possible' – needs emendation. As a first approximation, we might say that deliberative procedures are rational if and only if they are effectively directed to making one's life go as well as possible. (1998, p. 58, n. 5)

And in the text Gauthier goes on to say that in order for my deliberation to be rational, it must be "directed effectively at the realization of my concerns" (1998, p. 49). Gauthier concludes,

> The pragmatic standard that I embrace does not lead to the absurd view that rationality is simply a matter of what pays. (1998, p. 50)

In this way, Gauthier imposes a restriction on the pragmatic-effectiveness criterion, allowing it to operate only when another condition has been satisfied. This enables him to add what, in essence, we claimed earlier: "It is not valid to argue: *p*, because it would pay me to believe that *p*" (1998, p. 52).

We have seen that Gauthier regards the merits of a theory of rationality as dependent on its meeting two requirements: a directedness requirement, according to which a theory of rationality must be such that an agent's deliberation in accordance with it is *directed to* the furtherance of the agent's ends; and an effectiveness requirement, according to which a theory of rationality must be such that an agent's deliberation in accordance with it is *effective for* the furtherance of the agent's ends. More precisely, we can say that, according to Gauthier, the best theory of rationality is the one that satisfies the directedness requirement and satisfies the effectiveness requirement to as great an extent as does any theory of rationality satisfying the directedness requirement. And this refined approach seems to hold great promise. For both of its requirements are backed by strong intuitions. First, the effectiveness requirement has the support of the intuitions discussed in section 2 (intuitions about, e.g., the non-self-defeating character of rationality and the possibly of rational agents "stand[ing] as their own guarantors"). And the directedness requirement gives voice to the intuition that not all effective deliberation is rational; in order to be rational, deliberation must be (consciously) *directed* to the furtherance of one's ends.[14]

4.3 But does this refinement really dispose of all of the counterexamples to the unrestricted pragmatic-effectiveness criterion? The reply from counterexamples, it will be recalled (from subsection 4.1), is that cases such as that of the tyrant show that deliberative procedures that even Gauthier admits are patently irrational may possess, by the vagaries of circumstance, pragmatic effectiveness. In other words, the reply is that there are counterexamples—of which the tyrant case is just an extreme instance—that show the implausibility of the pragmatic-effectiveness criterion. Does Gauthier's refined criterion for evaluating theories of rationality answer this objection? To be sure, it takes care of some of the counterexamples: those in which effective deliberation is not directed at the furtherance of the agent's ends. But are all of the counterexamples like this? Are all of the cases in which we refuse to regard effective deliberation as rational deliberation such that our reason for so refusing is that the deliberation fails Gauthier's directedness requirement? Take, for example, the case of the

[14] It is important not to overstate Gauthier's refinement. According to Gauthier (2001), the directedness requirement is not to be taken as disqualifying, as irrational, those forms of deliberation in which the agent, while regarding the tyrant's orders as reasons for action, also sees how so regarding the tyrant's orders is a form of deliberation from which she benefits—even if she does not benefit from all, or even any, of the specific acts that such deliberation leads her to perform. The forms of deliberation that the directedness requirement *is* to be taken as disqualifying, as irrational, are those in which the agent regards the tyrant's orders as reasons for action and yet does not also see how so regarding the tyrant's orders is a form of deliberation from which she benefits. So the directedness requirement is (I think one might say) a requirement that the agent continually (by which I mean not at every moment, but at least in a permanent and ongoing way) *monitor* her deliberative practices with a view to ensuring that they—but, again, not necessarily the acts they lead her to perform—advance her aims.

The case of friendship shows the upshot of this understanding of the directedness requirement. The directedness requirement does not disqualify, as irrational, deliberation in which an agent takes herself to have reason to treat a person in a certain way *simply because she is that person's friend and regardless of whether so treating that person serves her (the agent's) interest*, as long as the agent also sees how so taking her friendship with that person, or friendships in general—i.e., sees how taking them as sources of reasons—makes her life go better (than it would if, e.g., she took only straightforwardly maximizing considerations as reasons for action). The directedness requirement does disqualify, as irrational, deliberation in which an agent takes considerations of friendship as reasons for action without monitoring the effect on her well-being of so taking such considerations. Unreflective friendship, like blind obedience to the tyrant, is the sort of deliberative disposition that violates the directedness requirement.

toxin puzzle. Even Gauthier's refined approach—effectiveness constrained by directedness—almost certainly selects, as best, a theory of rationality according to which it is rational for the agent to drink the toxin if he has won the money by intending to drink it (which is the only way in which he could have won the money without the story having broken down in some way). But many people would regard it as irrational for the agent to drink, since he then has nothing to gain. They may not *criticize* the agent for drinking, since they may not want to discourage agents from following through on their intentions when doing so is a net loss to them, since it's generally socially useful for people to be disposed to follow through on their intentions. But they would likely regard such conduct as irrational, strictly speaking.

Two replies—one misguided, the other somewhat more sensible—may be offered on Gauthier's behalf. First, it might be suggested that Gauthier could take care of such counterexamples by insisting that they, too, are cases in which effective deliberation is not appropriately "directed" and is, hence, not rational. But this reply, it should be obvious, would be the last thing Gauthier would want to say. For his purpose in qualifying the effectiveness requirement with the directedness requirement is to maintain the plausibility of an approach to evaluating theories of rationality on which a theory that endorses drinking the toxin—i.e., constrained maximization—can be judged best.

A second, and somewhat more sensible, reply is that counterexamples such as the case of the toxin puzzle simply beg the question against the pragmatic-effectiveness criterion and any theory that that criterion is used to defend. But is adducing such cases as counterexamples really an instance of begging the question? Or is it an instance of pointing out counterintuitive implications?

Certainly the fact that Gauthier himself adduces such cases as implications of his view does not render them ineligible to be called into service, by those who differ, as counterexamples. To see this, suppose for a moment that Gauthier refused to concede that effective deliberation in the realm of the tyrant would be irrational, and that we then adduced *that* case as a counterexample to his original, unqualified, assertion of the pragmatic-effectiveness criterion. Could not a defender of the pragmatic-effectiveness criterion say in that case, just as freely as in the toxin case, that adducing such a case as a counterexample is simply begging the question? There seems to be nothing about the toxin case that makes the "begging the question" rejoinder available there while precluding its use in the tyrant case. And so the fact that the "begging the question" rejoinder can be summoned to the defense of a position that Gauthier himself concedes the absurdity of should give us pause before we trust it as an adequate rejoinder to our claim that other cases, where directedness is not an issue, constitute counterexamples to the pragmatic-effectiveness criterion—even if they are cases that Gauthier cites as favorable implications of his approach.

I have argued that such counterexamples cannot simply be dismissed as question-begging. But, at the same time, I must concede that one man's counterexample is another man's favorable implication. And there does not seem to be any way of arguing for the rationality or irrationality of a particular case except by appealing to intuitions, theories, and criteria whose merits are very much the subject of debate. Since the full force of such considerations cannot be felt until after discussions of the criteria examined in the next two chapters, a comprehensive weighing of the competing considerations will be deferred until chapter VI. So the most that can be claimed at this point, on behalf of the reply from counterexamples, is not that it decisively disposes of the pragmatic-effectiveness criterion, but that it shows the possibility of certain cases' being regarded as compelling counterexamples against it.

5 The reply from circumstantial contingency

5.1 The reply from circumstantial contingency differs from the reply from counterexamples in two important ways. First, whereas the reply from counterexamples is easier to develop in the context of theories of rationality, this reply is easier to develop in the context of theories of morality. Second, whereas the reply from counterexamples was claimed only to achieve a stalemate against the pragmatic-effectiveness criterion, I offer the reply from circumstantial contingency as a conclusive refutation of the pragmatic-effectiveness criterion.

As we noted earlier (in subsection 2.6), the pragmatic-effectiveness criterion implies that the best theory of morality—the one such that deliberation and action in accordance with it are (morally) right—is the one such that its general acceptance in a society would advance that society's aims more than would the general acceptance of any other theory of morality. The reply from circumstantial contingency is predicated on some claims about the properties that would be possessed by the moral theory that satisfies this criterion. It then proceeds to show that the conception of morality implicit in the pragmatic-effectiveness criterion is an untenable one.

5.2 To see what sort of moral theory would be selected as the best moral theory by the pragmatic-effectiveness criterion, begin by considering why act consequentialism apparently would *not* be so selected. This theory says that the right act in any circumstance is the one that brings about results at least as good as those that any alternative act would bring about. Agents who subscribe to act consequentialism (i.e., agents who employ act consequentialism as their deliberative procedure) would seek, in every instance of deliberation and action, to follow this rule: Choose that act (or one of the acts, if there is a tie) that will have the best consequences. Now we saw in the last chapter that such agents can be expected to bring about outcomes worse than those brought about by other agents, such as those who subscribe to common-sense morality. It can be expected, we saw, that even if they do not have any psychological problems stemming from always aiming at overall happiness impartially conceived, and even if they work as a team in such a way as to steer clear of any deleterious interaction effects, then they will very likely be stymied by implementation problems ranging from time spent deliberating to miscalculating outcomes due to self-interested considerations' creeping into their deliberations.

This suggests that a theory of morality can be pragmatically effective only if it requires far less calculation on the part of agents than act consequentialism does. In order for this to be the case, a theory of morality would clearly need to consist of one or more rules that require much less calculation than the act-consequentialist rule does. But the system must not be too elaborate, or else it will be too cumbersome for agents to learn and retain. Brandt sets out the sources of this tension very clearly:

> What makes a moral code effective in a society? It must be suited to the level of intelligence and education of the society; so its application must not demand logical facility beyond the capacities of all but potentially good scientists or philosophers. It must provide in detail for problems of frequent occurrence in the society, without the necessity for long trains of inference. It must not contain items too numerous to be taught by the methods (e.g., classical conditioning) which must be used to interiorize moral principles; so it will probably restrict itself to matters of some importance in the society. (Special 'codes of ethics' for physicians and lawyers seem to do this: they speak directly of complex situations with which the relevant persons are often faced, and give direction for those which do not require deduction from abstract principle.) It looks as if a working moral code must comprise a set of specific directives like 'No cigar smoking!' These directives might not guide people to do exactly what ideally anyone would

> like them to do, but they are the best that can be done by the instrument of a moral code. (Brandt 1979, p. 181)[15]

These considerations, as suggested above, give rise to a tension that any set of rules aspiring to pragmatic effectiveness must withstand. As Baier writes, any such set of rules "must satisfy two opposing demands: that they be sufficiently general, or else they will not apply to a sufficient number of cases so that the individual has to master too many of them; and they must be sufficiently specific or else they will not provide the guidance they are supposed to give" (1995, p. 62). He goes on to develop this point in more depth:

> It should be obvious that the guidelines must strike a compromise between two opposing principles, the *principle of parsimony* and the *principle of specificity*. Since people should be able to recognize the relevance of facts without looking up a list of guidelines, such as a legal code book, there should be relatively few so that they can remember all or most and so recognize which of the facts of their case can be subsumed under the constituent of one or other of the familiar guidelines. To satisfy this first principle, the guidelines would have to be comparatively nonspecific, abstract, or general, rather than specific, concrete, or particular. But as we have already noted, the more general they are, the less helpful they will often be. (1995, p. 83)

Such considerations, then, will principally determine the character of the theory of morality that will prove to be maximally pragmatically effective.[16]

5.3 It seems, just as a matter of empirical psychological fact, that the moral theory that would prove to be maximally pragmatically effective will have to be rather limited in its complexity. But now remember how the pragmatic-effectiveness criterion works: it selects, as the best moral theory—the one such that deliberation and action in accordance with *it* are right—whatever theory is maximally pragmatically effective. It follows that the pragmatic-effectiveness criterion implies that the best moral theory sets forth an account of morality—an account of right conduct, duties, and so on—that is rather limited in its complexity. Moreover, precisely how limited in its complexity it is will be a function of the mental capabilities of the agents in the society in question.

But this is where a problem arises. For it is implausible to regard the content of morality—which is what we are really talking about—as restricted in this particular way. To see this, imagine a society in which agents almost never find themselves outside of situations that cannot be covered by a handful of very simple rules. Moreover, the agents in this society are

[15] See also Brandt (1979, p. 232 and pp. 273–274).

[16] On this point see also Hare (1981, pp. 35–39), Kagan (1989, p. 37), and Crisp (p. 107).

remarkably inept when it comes to assessing the likely consequences of their actions. Suppose, further, that as a result of these and any other necessary considerations, the optimal moral code in this society—the most pragmatically effective moral theory for it—is one with no rules that leave much to agents' discretion. In particular, the optimal moral code does not include a disaster-avoidance provision: a rule lexically prior to all other rules, enjoining agents to act so as to prevent disasters (where a disaster is understood to be an outcome *much* worse than some alternative) when they can do so without imposing great costs on themselves or others. The agents in this society so seldom find themselves in situations where such a rule would be genuinely applicable, and would so often misapply it (to the detriment of the values embodied in the other rules, because they would sometimes be overridden), that the optimal code just does not include any such provision.

Now suppose that I am an agent in this society, and suppose that I find myself in a situation in which I can prevent a disaster without any great loss to myself, to others, or to the values embodied in the optimal rules. But suppose there will be a slight cost in terms of the value embodied in one of the other rules, so that the moral code actually prohibits me from doing what would prevent the disaster. In regard to such a case it seems wholly unreasonable to say, as the pragmatic-effectiveness criterion does, that I am forbidden from preventing the disaster because a rule enjoining such conduct is not part of my society's optimal code. Indeed to remove any possible sources of confusion, let us assume not only that a certain act of mine will prevent a disaster, but that I *know* this: that is, not only is it true, but my belief that it is true is reasonable and justifiable. We may even stipulate that I am aware of the possible consequences of my act in terms of setting an example (of breaking the society's optimal rules) that it would be dangerous for others to follow, and that I accurately and justifiably believe that even when this and all other costs of my act are taken into account, the outcome is still *much* worse if I do not act than if I do. In such a case, I submit, it is simply implausible to say that I am forbidden from preventing the disaster.

The general problem that this example illustrates is that the pragmatic-effectiveness criterion makes the content of morality for a society depend overmuch on the mental capabilities of the agents in that society. The problem is not that it makes a particular agent's duty in a particular case dependent on the mental capabilities of that agent; indeed it is perfectly reasonable to regard a particular agent's duties as conditioned in some way by his mental capabilities. Rather, as Kagan writes, the pragmatic-effectiveness approach goes astray because it

> assumes that once we know what set of norms it would be best to have taught in a society we also know the complete set of moral considerations relevant to an individual agent's actions; morality is exhaustively captured by the optimal set of social norms. But this is a substantive, controversial thesis, and it is hard to see what reason there could be to accept it. (1989, p. 37)

To be fair, we should take care not to overstate our case: the pragmatic-effectiveness criterion need not be a substantive thesis; it might be advanced as an analytical one, about the concept of morality itself. But however it is construed, we have seen that it is an untenable approach to evaluating moral theories.

5.4 To bring out this conclusion more vividly, let us elaborate the case we have been discussing. Retain all of the foregoing suppositions, and add this: a generation or two from now, a disaster-avoidance provision will be part of the society's optimal code. (This may happen because the occasions of possible disaster-avoidance become more frequent, or because agents' judgment improves, or for any of many other possible reasons.) And suppose that, once this provision becomes part of my society's optimal code, I find myself in a situation identical in all relevant respects to the earlier one, in which I was not allowed to prevent the disaster. Does it make sense to say that now I *am* allowed to prevent the disaster? Note that the only way the circumstances have changed is that I now live in a society in which a different set of rules is optimal from before. But the details of the particular case are exactly the same: the outcomes in question are the same, the cost in terms of other values would be the same, and so on. Indeed this is simply what's meant by saying that the situation in which I now find myself is the same in all relevant respects as the one I was in before. Now surely what an agent ought to do in one case must match what the agent ought to do in another case that is identical in all relevant respects. And the pragmatic-effectiveness criterion's implication that my duties are different in the two cases must therefore be rejected.

The foregoing argument rests on the uncontroversial claim that what an agent ought to do in one case must match what the agent ought to do in another case that is identical in all relevant respects. But it also rests on the claim that the cases under discussion are identical in all relevant respects, and this may be contested. For it may be claimed that the stipulated difference in the society's optimal rules constitutes a relevant difference between the two cases. But it is really reasonable to say that this difference—a difference that has no manifestation in the cases under discussion—still counts as a relevant difference between the two cases? It seems unreasonable to say, as the pragmatic-effectiveness criterion requires, that two cases that differ only in *this* respect differ in some *relevant* respect. It follows that the pragmatic-effectiveness criterion either (1) cannot accommodate the uncontroversial claim that what an agent ought to do in one case must match what the agent ought to do in another case that is identical in all relevant respects or (2) is committed to a very dubious substantive thesis about what counts as a relevant difference between two cases.

5.5 There is one other implication of the case that we have been discussing that should be pointed out. I have been dwelling on the fact that on the pragmatic-effectiveness criterion, my duties change from one case to the next. But my duties change from one case to the next only because the moral theory that the pragmatic-effectiveness criterion selects as the best one changes from one period to the next. That is, according to the pragmatic-effectiveness criterion, there is no single moral theory that is best in some atemporal way; rather, 'the best moral theory' is a designator that ranges freely over the indefinitely many moral theories that can be imagined, picking out at any moment that theory that is optimally suited to the present circumstances. And as agents' mental capabilities and the kinds of situations in which they find themselves constantly vary, so does—according to the pragmatic-effectiveness criterion—the content of morality. On the pragmatic-effectiveness criterion, then, morality has no stable content; its content varies continuously, always answering to the circumstances that prevail at the moment.

But should we conceive of morality in this way? Or should we suppose that, although any reasonable moral theory will be responsive to the circumstances, the best moral theory is one that persists through time and accommodates changes in circumstances, rather than being displaced by them in order to make room for the *next* "best moral theory"? It turns out, then, that we can accept the pragmatic-effectiveness criterion only at the cost of allowing that no society that changes through time can have a *best* moral theory that persists through time.

5.6 I have been arguing that the pragmatic-effectiveness criterion, by making the content of morality dependent on the contingencies of the circumstances in question, is committed to several implausible implications: that it makes the content of morality implausibly dependent on the mental capabilities of the agents in the society in question (subsection 5.3), that it either denies that similar cases must be judged similarly or involves dubious substantive claims about what causes cases to be dissimilar (subsection 5.4), and that it suggests that the best moral theory is constantly shifting in its content (subsection 5.6).

If we reject the pragmatic-effectiveness criterion, then we must allow that conduct in accordance with a society's optimal rules may be wrong. But it is perfectly reasonable to insist on this possible divergence between the best rules and right acts. As Hare writes,

> The winner of a game of backgammon is the player who first bears off all his pieces in accordance with the rules of the game, not the one who follows the best strategies. Similarly in morals, the principles which we have to follow if we are to give ourselves the best chance of acting rightly are not definitive of 'the right act'. (1981, p. 38)

Rejecting the pragmatic-effectiveness criterion, then, saves one from several implausible implications about the nature of morality while leaving one with a perfectly plausible understanding of how right conduct is related to, but not necessarily identical to, conduct that accords with the most pragmatically effective decision procedure.

6 Some rivals of straightforwardly maximizing theories

6.1 The last three sections were devoted to three replies that might be offered in order to dispute the significance of the self-defeat of straightforwardly maximizing theories. In this section, I shall bracket the question of the significance of the self-defeat of those theories and shall revisit the question of just how self-defeating or pragmatically ineffective such theories are, relative to certain of their rivals. I shall argue, in particular, that common-sense morality and rule-based theories, such as rule egoism and rule consequentialism, are far less pragmatically effective than one might initially think.

6.2 It is often thought that common-sense morality, being the product of more reflection and refinement informed by practice than any other moral theory, must score exceptionally well on the scale of pragmatic efficacy. But as Parfit argues, the morality of common sense is deficient in certain ways. Parfit begins by pointing out the following:

> Most of us hold that there are certain people to whom we have special obligations. These are the people to whom we stand in certain relations—such as our children, parents, friends, benefactors, pupils, patients, clients, colleagues, members of our own trade union, those whom we represent, or our fellow-citizens. We believe that we ought to try to save these people from certain kinds of harm, and ought to try to give them certain kinds of benefit. Common-Sense Morality largely consists in such obligations. (1984, p. 95)

Parfit goes on to concede that common-sense morality leaves some room for a general duty to help strangers. But the point stands that common-sense morality instructs us not to be impartial benefactors of mankind: "I ought to save my child from some harm rather than saving a stranger from a *somewhat* greater harm. My duty to my child is not overridden whenever I could do somewhat greater good elsewhere" (Parfit 1984, p. 95).

These special obligations may cause agents occupying certain roles, such as parents, to find themselves in dilemmas resembling prisoner's dilemmas. To see this, suppose that parents of two unrelated children find themselves in a situation in which each child is at risk of suffering a greater or a lesser harm, or possibly both. Suppose also that each parent can either save his own child from the lesser harm or the other's child from the greater harm. Now if the difference between the greater harm and the lesser harm is large enough, then common-sense morality would direct each parent to save the other's child (from the greater harm), out of a general duty to help others avoid suffering great harms. Even though each parent would then be declining to shield his own child from the lesser harm, this is the preferred outcome: since each parent can only prevent one of the harms, the best they can hope for is for each to prevent the greater harm to the other's child. (Conceivably, a parent could save his own child from the lesser harm and hope for the other parent to save him from the greater, but this outcome—though, like the preferred outcome, Pareto optimal—can hardly be recommended.) But given the special obligations that common-sense morality includes, there must be some difference between the greater harm and the lesser harm that is small enough so that the following would be true: common-sense morality directs each parent to save his own child from the lesser harm, even though that means that each child will thereby suffer the greater harm. Such cases show how common-sense morality may lead to Pareto-inferior outcomes (Parfit 1984, p. 96).

Moreover, cases such as the one outlined by Parfit may arise naturally out of any case that has the structure of the prisoner's dilemma. To see this, consider the prisoner's dilemma itself, and suppose that not the suspects, but their lawyers, are the agents making the choices. Then, if each lawyer has a special obligation to his client, they find themselves in what Parfit calls the prisoner's lawyer's dilemma: "If both lawyers give priority to their own clients, this will be worse for both clients than if neither does" (1984, p. 98). In the same way, any dilemma with the form of the prisoner's dilemma can yield something like the prisoner's lawyer's dilemma:

> Any self-interested Dilemma may thus yield a moral Dilemma. If one group face the former, another group may in consequence face the latter. This may be so if each member of the second group ought to give priority to some members of the first. (1984, p. 98)

Now the prisoner's dilemma itself may not arise very often outside of police dramas on television, but cases that have that structure—what Parfit calls self-interested dilemmas—are not uncommon. And the kinds of special obligations that common-sense morality imposes are numerous and diverse. As a result, there should be many cases in which the special obligations imposed on agents by common-sense morality may lead them to act so as to bring about sub-optimal outcomes.[17] As a result, it can hardly be assumed that common-sense morality is pragmatically optimal.

6.3 Rule-based theories are ones that refer explicitly to rules, or other guidelines for making choices, that are optimal in some sense: whether optimal when adopted by everyone or when adopted by just one person.[18] A standard form of rule consequentialism, for example, says that

> An act is right if any only if it would be allowed by rules whose general acceptance would have better consequences than the general acceptance of any other rules.

Hooker advances a theory very much like this one (2000, p. 32). But rules are not the only guidelines for making choices that what I call rule-based theories refer to. They also refer to such guidelines as policies, commitments, motives, and consciences.

Kupperman, for example, claims that "the form of consequentialism that we ought to take most seriously" is not act consequentialism, but the following: "In choosing among the policies and commitments that govern our moral lives, we should choose those that are likely to have the best consequences" (1980, p. 327).[19] Although this view is not itself a moral theory, since it does

[17] Parfit's argument is criticized by Setiya. But Setiya's criticism depends on the assumption that it is important, according to common-sense morality, to act morally, above and beyond acting to bring about certain results. And this assumption may be false. Moreover, even it were true, Setiya would have shown only that Parfit's argument does not show that common-sense morality is self-defeating in a particularly strict sense of that term (one, admittedly, that Parfit uses). He would not, I think, have given us reason to doubt that common-sense morality is self-defeating in the sense of being pragmatically ineffective.

[18] Note that they do not include theories such as constrained maximization and resolute choice. For instead of referring to optimal rules without specifying their content, each of these theories outlines a specific decision procedure that, in addition, is claimed to be pragmatically effective.

[19] Kupperman also says that at "at a *minimum* consequentialism includes the recommendation to judge moral policies and commitments on the basis of consequences plus the recommendation to choose directly in some cases that alternative likely to have the best consequences" (1980, p. 328). If this statement is supposed to be consistent with the one quoted in the text, then Kupperman must be supposing that the policies and commitments

not claim that certain *acts* are right and others are wrong, we can easily imagine a proponent of this view defending the moral theory that

> An act is right if and only if it would be permitted by the policies and commitments whose general acceptance would have the best consequences.

This theory, though referring to policies and commitments rather than rules, has obvious similarities to standard forms of rule consequentialism.

Another sort of rule-based theory is offered by Adams. His theory revolves around motives:

> one pattern of motivation is morally better than another to the extent that the former has more utility than the latter. The morally perfect person, on this view, would have the most useful desires, and have them in the most useful strengths; he or she would have the most useful among the patterns of motivation that are causally possible for human beings. (p. 470)

Again, this view is not itself a moral theory, since it is not about acts but about patterns of motivation and agents. But Adams adds that "a very natural position for a motive utilitarian to take in the ethics of actions" is a view that may be called "conscience utilitarianism":

> we have a moral duty to do an act, if and only if it would be demanded of us by the most useful kind of conscience we could have. (p. 479)

This view, since it is about acts, counts (for us) as a moral theory. And Brandt defends his own, very similar, version of conscience utilitarianism (1996, p. 145).[20]

6.4 It is not uncommon, then, to find authors making explicit reference to the rules, policies, commitments, motives, or consciences that would be pragmatically optimal. And it might seem that some such theory would easily satisfy the pragmatic-effectiveness criterion. But let us consider one of these theories in more depth. Let us focus, as a representative example, on Kupperman's theory, which refers to the optimal "policies and commitments." Now consider the adoption and acceptance of Kupperman's theory by one or more agents. That is, one or more agents come to believe, and continue to believe, that an act is right if and only if it would be permitted by the policies and commitments whose general acceptance would have the best consequences. The crucial thing to realize at this point is that when we hypothesize the adoption

that are likely to have the best consequences include provisions for choosing in a consciously optimizing way in some cases. Such a supposition, though, is an empirical one—not part of the analytical content of consequentialism. This issue also arises in Brandt's work (1967, pp. 58–59).

[20] Although Hooker's theory is stated in his text as a form of rule consequentialism (2000, p. 32), his remarks on how it is to be interpreted, later in his book, indicate that he really has in mind a form of conscience consequentialism (2000, pp. 89–92).

and acceptance of this theory by an agent—as we need to do in order to ascertain this theory's pragmatic effectiveness—we do not thereby hypothesize the adoption and acceptance by that agent of the optimal policies and commitments. Rather, we only hypothesize the adoption and acceptance by the agent of the belief that those optimal policies and commitments—whatever their specific content may be—are determinative of right conduct. Then, it is up to the agent to figure out what those optimal policies and commitments are.

Of course, presumably the agent needn't ever figure out the entirety of what those policies and commitments are. For he will need to figure out their content at all only in order to ascertain what acts are right and wrong in particular cases, and ascertaining what acts are right and wrong in particular cases may require the agent only to figure out the parts of the optimal policies and commitments that are relevant to the case at hand. But as the agent faces situations of different kinds, then he will need to be engaged in an ongoing project of filling out his best guess as to what the optimal policies and commitments are. Of course, he may learn from experience and employ rules of thumb to guide his thinking, but it will always be up to him to figure out (or to regard himself as having already figured out) what the optimal policies and commitments are.

So a theory of the kind we are now considering directs any agent who subscribes to it to answer a certain question as a means to making any moral decision he confronts. That question is, "What are the optimal moral policies and commitments?" (or, perhaps, "What are the optimal policies and commitments for a case of the sort you now face?"). The agent must do the work of answering the question on his own—or, at least, not with the help of the theory. In short, hypothesizing that the agent adopts and accepts the theory means hypothesizing only that the agent *asks a certain question* in order to decide how to act, not that he already has the *answer* to it.

Now it might be thought that answering this question would not be very hard—agents would tend to regard the rules of common-sense morality as the ones that it's best to follow. Hodgson, for example, claims this (pp. 67–70). But once they realize the possibility of prisoner's lawyers dilemmas, will they continue to regard those rules as the pragmatically optimal ones? Or will they try to formulate some revised version of common-sense morality that removes those defects? So it cannot be assumed that there is an easy answer available to agents. On the contrary, the optimization problem that theories such as rule consequentialism put to them is a very complicated one.

6.5 To understand the nature of the phenomenon being discussed, let us distinguish it from another phenomenon with which it may be confused. Brandt writes the following:

> A natural mistake about the content of the 'ideal' code is to assume that the welfare-maximizing code which rational people would support would be a one-principle code after all, like act utilitarianism, just one principle: 'Do whatever would be required by the moral code the currency of which would maximize welfare.' Is this the whole moral code which rational people would support? The answer to the question is: 'No, rational persons would not support that one-principle moral code.' We can see why not if we try to spell it out. Suppose it were true that the moral code the currency of which would maximize welfare is just this one-principle code. Then what we should have to teach is: 'Do whatever would be required by the moral code, "Do what would be required by the moral code . . ." ' and so on indefinitely. In other words: 'Follow the principle which is to follow the principle. . . .' Either there is some specifiable set of rules the currency of which would maximize welfare, or there is not. If there is not, then the one-principle [code] becomes, in the phrase of C. I. Lewis, a 'perpetual stutter'. If there is, then it is the set of principles rational persons would want taught, and is the moral code the currency of which would maximize welfare. (1979, pp. 294–95)

Brandt is right to say that the one-principle code he specifies is not "the welfare-maximizing [i.e., maximally pragmatically effective] code which rational people would support." But he is wrong to say that the reason for this is that the code becomes a "perpetual stutter." For if, as Brandt asks us to suppose, "it were true that the moral code the currency of which would maximize welfare is just this one-principle code"—i.e., if it were true that the one-principle code were pragmatically optimal—then what we would have to teach is not the unfinishable sentence that Brandt suggests, but simply that one-principle code itself: 'Do whatever would be required by the moral code the currency of which would maximize welfare." Why would this be what we would have to teach? Well, for the simple reason that we have hypothesized that it is the pragmatically optimal moral code. To hypothesize that it is the pragmatically optimal moral code just is to hypothesize that it, and not some complicated variant of it, is what we would have to teach. Of course, one oddity about such a situation would be that the code we would be teaching would refer to itself, and it may seem doubtful that a code of such abstraction and self-reference could be the pragmatically optimal one. But that is just my point: that such a code would fail to be pragmatically optimal for *this* reason, not because of any sort of "perpetual stutter" problem.

In a more recent work, Brandt corrects his earlier apparent misdiagnosis of the malady that besets rule-based theories. He writes,

> the conscience-utilitarian *theory* does not include a specification of moral motivations it would be best to have prevalent and taught. . . . The general theory is essentially a directive for *thinking* to be done: things a person must do to identify the moral codes it would be best to have prevalent and taught, all costs and benefits considered. (1996, pp. 154–155)

This realization, better than the concerns he had earlier about a "perpetual stutter," ground his conclusion that

> there is strong reason to think that the currency of a carefully selected plural code [i.e., a code with several rules] in a society will produce considerably more welfare or happiness, on balance, than any of the one-principle codes that have historically been advocated. (1979, pp. 295–96)

So, theories such as rule consequentialism stand little chance of being pragmatically optimal.

6.6 We have seen, then, that theories such as rule consequentialism—although seeming to be plausible contenders for pragmatic optimality because of their explicit reference to optimal rules, policies, and so on—will almost definitely score low in terms of pragmatic effectiveness, because of the difficulty of the task they set for agents: the task of figuring out what those optimal rules and policies are. It follows that the pragmatic-effectiveness criterion provides no support for rule consequentialism. A similar argument would apply, of course, to rule egoism.

Now it may be noticed that the sources of pragmatic ineffectiveness for rule-based theories just discussed are of the first of the three kinds distinguished in chapter II: problems of implementation, rather than problems of human desires and theoretical demands or problems of dynamic inconsistency. And the theoretical importance of this kind of source of self-defeat might be doubted: it might be thought that problems of implementation should be bracketed as indicative more of the accidents of human capabilities than of the intrinsic properties of normative theories. If this claim could be substantiated, then this section's conclusions about rule-based theories (though not those about common-sense morality) could be set aside as not really bringing to light anything of theoretical interest about those theories.

But this claim, here offered on behalf of rule-based theories, is essentially the same as a claim we considered in chapter II on behalf of act-based theories. That claim—given forceful expression by Mill's remark that "There is no difficulty in proving any ethical standard to work ill, if we suppose universal idiocy to be conjoined with it" (quoted in subsection 2.4 of chapter

II)—was answered by pointing out that not every case of pragmatic ineffectiveness due to implementation problems requires us to suppose universal idiocy. On the contrary, it is sufficient if we simply suppose the capacity for means-end reasoning and resisting temptation that normal human beings have. And surely this is not an unreasonable environment to postulate for the implementation of a normative theory. For the limitations inherent in such capacities are not accidental limitations due to humanity's current state of technological progress, as is (perhaps) the fact that it presently takes at least a half a day or so for a person to travel to the other side of the world; not are they even accidental limitations of our physical world, as is the fact that it takes more than a twentieth of a second for any sort of signal to travel to the other side of the world. Rather, these limitations are inherent in the very nature of an organism with finite capacities for the acquisition, storage, recall, and manipulation of data. That is, they are inescapable features of the human condition. Of course, it might be replied that in focusing on the *human* condition we set our sights too low, but that is a debate for another day, and another dissertation.

7 Conclusion

In this chapter I have endeavored to assess the significance of the results of the previous chapter: to ascertain to what extent the pragmatic ineffectiveness of straightforwardly maximizing theories, as shown in the previous chapter, should be regarded as a reason to set them aside and to opt instead for certain of their indirectly maximizing rivals. After identifying an argument for the pragmatic-effectiveness criterion (section 2), I discussed three replies to it. The reply from self-effacement was seen to be unsuccessful (section 3), the reply from counterexamples was claimed only to achieve a stalemate against that criterion (section 4), and the reply from circumstantial contingency was claimed to decisively dispose of that criterion (section 5). Finally, it was shown that the possibility of prisoner's lawyer's dilemmas casts doubt on the pragmatic optimality of common-sense morality and that the excessive abstractness of rule-based theories such as rule consequentialism and rule egoism results in their being beset with the same sort of pragmatic ineffectiveness that plagues straightforwardly maximizing theories for which they are offered as substitutes (section 6).

IV

Dispensing with the Publicity Condition

> It would be natural to want the best theory . . . not to be self-effacing. If the best theory was self-effacing, telling us to believe some other theory, the truth . . . would be depressingly convoluted. It is natural to hope that the truth is simpler: that the best theory would tell us to believe itself. But can this be more than a hope? Can we assume that the truth *must* be simpler? We cannot.
>
> —Derek Parfit (1984), p. 24

1 Introduction

We saw in the last chapter that if a normative theory is charged with being self-defeating in the sense of being pragmatically ineffective, then its defenders might reply by claiming that their theory is not self-defeating because it is, rather, self-effacing: it saves each agent who subscribes to it from the non-optimal outcome to which it (being pragmatically ineffective) would otherwise consign the agent by enjoining him to adopt, as a decision procedure, whatever theory whose adoption by him would be optimal—even though that theory will typically be a rival of the one in question. And we saw that this reply essentially misses the point of the pragmatic-effectiveness criterion, and so does not constitute an adequate reply to criticisms based on that criterion.

But this does not put the issue of self-effacement to rest. For opponents of self-effacing theories tend to find, in this feature of them, the basis for a different objection, quite unrelated to issues of pragmatic effectiveness, to the effect that they violate a requirement commonly known as the *publicity* condition. Moreover, the defender of a self-effacing theory cannot dodge this charge by declining to appeal to the self-effacing character of his theory in response to the pragmatic-ineffectiveness charge, for whether a theory is self-effacing is not a question that is settled by the wishes of any of its defenders. Of course, it is theoretically possible for someone to construct a variant of a self-effacing theory that is not self-effacing, but we shall see later in this chapter that such variants would be unsatisfactory on other grounds. With this promissory note having been issued, I propose to proceed on the assumption that the self-effacement objection is one that defenders of certain theories must face, not one that they can dodge by declining to appeal to self-effacement in their reply to the self-defeat objection or by advocating non-self-effacing variants of their theories.

The self-effacement objection is one of a set of objections that correspond to different forms of the publicity condition, and my aim in this chapter is to show that in none of its forms is the publicity condition a reasonable requirement to impose on normative theories. In order to defend this claim, I shall begin, in section 2, by formulating and distinguishing three versions of the publicity condition and by reviewing some of the admissions of self-effacement that are found in the literature on consequentialist theories, as well as some evidence of the validity that the publicity condition is widely thought to have. That will set the stage for an investigation, in section 3, into some of the considerations that may seem to justify the publicity condition and, in section 4, for the presentation of two independently sufficient refutations of it. I shall offer some concluding reflections in section 5.

2 Versions and violations of the publicity condition

2.1 Having only gestured at the general idea of the publicity condition in the introductory section, I endeavor in this section to formulate and to distinguish three versions of this requirement. We can approach the task of formulating these distinct versions by imagining some of the ways in which a normative theory may run afoul of the general idea of the publicity condition. A normative theory runs afoul of the general idea of the publicity condition in a particularly flagrant way if the theory implies, in certain circumstances, that *every* agent in the group to which it applies (such as a particular society, or all rational creatures) ought not to subscribe to it. Borrowing a term found in the passage from Parfit's *Reasons and Persons* that I have chosen as the epigraph for this chapter, let us call such theories *self-effacing*. Then one version of the publicity condition may be formulated as follows:

> *The ban on self-effacing theories*: A normative theory is unacceptable if circumstances may arise in which it requires every agent in the group to which it applies not to subscribe to it.

So, to satisfy this requirement, all a theory needs to do is always (that is, in all circumstances) allow some agent or agents in the group to subscribe to it—even if it also sometimes or always (that is, in some circumstances or all circumstances) implies that some of those agents ought *not* to subscribe to it. If a normative theory violates this requirement, it runs afoul of the general idea of the publicity condition in a particularly flagrant way.

But there are other ways in which a normative theory may run afoul of the general idea of the publicity condition. For example, even if a normative theory does not require *every* agent in the group to which it applies not to subscribe to it, it may still be thought to run afoul of the general idea of the publicity condition in some way if it ever requires even *some* agents not to subscribe to it. Borrowing a term from Sidgwick, let us say that such theories are *esoteric*. This suggests another, more demanding, version of the publicity condition, which can be formulated by replacing the word 'every' in the ban on self-effacing theories with the word 'some':

> *The ban on esoteric theories*: A normative theory is unacceptable if circumstances may arise in which it requires some agent in the group to which it applies not to subscribe to it.

To satisfy this requirement, then, a theory must always (that is, in all circumstances) allow *every* agent in the group to subscribe to it. This requirement is more demanding than the ban on self-effacing theories, a fact brought out particularly plainly when one notes that self-effacing theories form a subset of esoteric theories: some esoteric theories are self-effacing, while the rest are, we might say, only partially self-effacing (which of course does not count as being self-effacing *simpliciter*).

Perhaps no recent author is more responsible for drawing attention to the publicity condition—indeed, for causing it to be referred to in this way—than John Rawls. And it might be thought that the publicity condition as formulated by Rawls is equivalent to either the ban on self-effacing theories or the ban on esoteric theories. But although Rawls does imply an endorsement of the ban on esoteric theories, he does so in his discussion of what he calls the universality condition. There he writes that

> principles are to be universal in application. They must hold for everyone in virtue of their being moral persons. Thus I assume that each can understand these principles and use them in his deliberations. This imposes an upper bound of sorts on how complex they can be, and on the kinds and number of distinctions they draw. Moreover, a principle is ruled out if it would be self-contradictory, or self-defeating, for everyone to act upon it. Similarly, should a principle be reasonable to follow only when others conform to a different one, it is also inadmissible.

> Principles are to be chosen in view of the consequences of everyone's complying with them.[1] (1999b, p. 114)

What he conceives of as the publicity condition goes further:

> A third condition [after generality and universality] is that of publicity, which arises naturally from a contractarian standpoint. The parties assume that they are choosing principles for a public conception of justice. They suppose that everyone will know about these principles all that he would know if their acceptance were the result of an agreement. Thus the general awareness of their universal acceptance should have desirable effects and support the stability of social cooperation. The difference between this condition and that of universality [which, as just noted, implies the ban on esoteric theories] is that the latter leads one to assess principles on the basis of their being intelligently and regularly followed by everyone. But it is possible that all should understand and follow a principle and yet this fact not be widely known or explicitly recognized. (1999b, p. 115)

This is the last, and the most demanding, of the versions of the publicity condition that we shall set out. The distinguishing feature of this condition, which I shall refer to as *Rawls's publicity condition*, is expressed in Parfit's observation that it says of a theory that "it must be a theory that everyone ought to accept, *and publicly acknowledge to each other*" (1984, p. 43, emphasis added).[2]

As Rawls and Parfit indicate, Rawls's publicity condition requires more of a theory than that it not be esoteric (not to mention self-effacing). For a theory could satisfy the first two versions of the publicity condition—by allowing that every agent may, or even ought to, subscribe to it—and still be such that, if everyone accepted it in certain circumstances, general awareness of this fact would have undesirable effects (such as compromising the stability of social cooperation). Such a theory, if universally accepted in those circumstances, might require some agents to take certain steps to keep this universal acceptance from becoming generally

[1] This last sentence, in conjunction with Rawls's insistence (stated earlier in the block of quoted text) that the principles be such that "each [agent] can understand these principles and use them in his deliberations," suggests that Rawls endorses a pragmatic-effectiveness criterion, of the kind discussed in the previous chapter, for the evaluation of a moral theory. But this is not the case. Considerations of pragmatic effectiveness are regarded as reasons only by the (imaginary) choosers in the original position; and the original position, in turn, is defined, in part, by constraints (such as the veil of ignorance) that are based on intuitions (about, e.g., fairness) that are not reducible to concerns about pragmatic effectiveness.

[2] Similarly, Hare writes that "the so-called publicity requirement by which Rawls and others set store" requires of a moral theory "that it could be *openly avowed* without defeating its object" (1997, p. 124, emphasis added).

The forgoing account of Rawls's publicity condition is based on *A Theory of Justice*. In his later "Kantian Constructivism in Moral Theory," Rawls characterizes his version of the publicity condition in still more detail. See Rawls (1980, pp. 324–327).

known (or, if it is already generally known, to make it not generally known), while not requiring any agent to interfere with that universal acceptance per se. Of course, one way of keeping the universal acceptance of the theory from becoming generally known is to keep the universal acceptance of the theory from being a fact (that is, keep the theory from being universally accepted), and if the theory prescribed *this* as the means of avoiding that undesirable general awareness, then the theory would be esoteric and possibly self-effacing, too. But a theory doesn't *have* to be esoteric or, especially, self-effacing, in order to violate Rawls's publicity condition.

So Rawls's publicity condition is the most demanding of the three versions of the publicity condition specified here, with the ban on esoteric theories being less demanding and the ban on self-effacing theories being the least demanding. We noted earlier that self-effacing theories form a subset of esoteric theories; those, in turn, form a subset of theories that violate Rawls's publicity condition. Although these relationships may be straightforward enough not to need a diagram in order to be fully understood, it will be convenient later to be able to refer to the following figure.

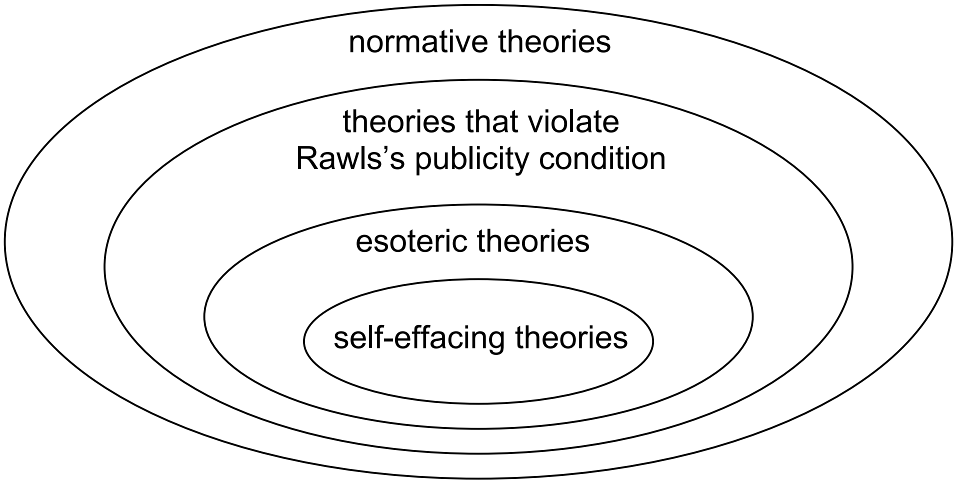

Figure IV.1

So in its least demanding version the publicity condition rejects only those theories in the innermost oval (the one for self-effacing theories); in an intermediate version it rejects all theories in the 'esoteric theories' oval, and in its most demanding version—Rawls's publicity condition—it rejects all theories except those outside the next-to-largest oval.

These three versions of the publicity condition, being of varying degrees of demandingness, are plainly distinct in the context of moral theories. But are they distinct in the context of theories of rationality? One structural way in which theories of rationality differ from theories of morality is that the "group" to which a theory of rationality pertains is just one person, in the sense that a theory of rationality applies to one person at a time, whereas a theory of morality applies to multiple persons—typically everyone in a certain society, if not all rational or sentient creatures—at a time. With this structural feature of theories of rationality in view, the possibility of a theory of rationality's being esoteric but not self-effacing seems doubtful. For if a theory of rationality is esoteric, then there is at least one person in the relevant group who (according to the theory) ought not to subscribe to it; but because the relevant group contains only one person, the theory implies that *everyone* in the relevant group ought not to subscribe to it. So any esoteric theory of rationality is a self-effacing one. What about Rawls's publicity condition? The possibility of an agent's subscribing to a theory (or, in Rawls's terms, "intelligently and regularly follow[ing]" or "understand[ing] and follow[ing]" it) but not being aware of this fact about his mental state also seems doubtful. And if the only way for an agent not to be aware of his subscribing to a theory is for him not to subscribe to it, then a theory of rationality failing Rawls's publicity condition is also self-effacing. So in the context of theories of rationality, the three versions of the publicity condition collapse into one; whereas in the context of moral theories, Rawls's publicity condition is more demanding than the ban on esoteric theories, and the ban on esoteric theories is more demanding than the ban on self-effacing theories.

2.2 Although the versions of the publicity condition just specified apply, in principle, to normative theories of all kinds, the incompatibility of consequentialist moral theories with requirements resembling these has long been a leading concern of both advocates and critics of versions of this theory. The *locus classicus* on this topic is Sidgwick's *The Methods of Ethics*, in which utilitarianism's violation of the publicity condition is described so explicitly, and in such thorough anticipation of later discussions, that passages from that book are worth quoting at some length. Reflecting on the likely pragmatic ineffectiveness of his utilitarian version of act consequentialism, Sidgwick writes that

> on Utilitarian principles, it may be right to do and privately recommend, under certain circumstances, what it would not be right to advocate openly; it may be right to teach openly to one set of persons what it would be wrong to teach to others; it may conceivably be right to do, if it can be done with comparative secrecy, what it would be wrong to do in the face of the world; and even, if

> perfect secrecy can be reasonably expected, what it would be wrong to recommend by private advice or example. (p. 489)

Sidgwick acknowledges the "paradoxical character" of these conclusions (p. 489), adding that

> there is no doubt that the moral consciousness of a plain man broadly repudiates the general notion of an esoteric morality, differing from that popularly taught; and it would be commonly agreed that an action which would be bad if done openly is not rendered good by secrecy. (pp. 489–490)

But rather than admitting or denying the soundness of either of these common opinions—that the notion of an esoteric morality is to be repudiated and that an act which would be bad if done openly is not rendered good by secrecy—Sidgwick points out that the latter opinion is one that "there are strong utilitarian reasons for maintaining generally," mainly (but not only) because

> it is obviously advantageous, generally speaking, that acts which it is expedient to repress by social disapprobation should become known, as otherwise the disapprobation cannot operate; so that it seems inexpedient to support by any moral encouragement the natural disposition of men in general to conceal their wrong doings. (p. 490)[3]

Thus, Sidgwick, rather than treating "the moral consciousness of a plain man" as a check on the acceptability of utilitarianism and its conclusions, simply embraces them as features of the world that a utilitarian would seek to maintain. Sidgwick concludes as follows:

> The opinion that secrecy may render an action right which would not otherwise be so should itself be kept comparatively secret; and similarly it seems expedient that the doctrine that esoteric morality is expedient should itself be kept esoteric. Or if this concealment be difficult to maintain, it may be desirable that Common Sense should repudiate the doctrines which it is expedient to confine to an enlightened few. And so a Utilitarian may reasonably desire, on Utilitarian principles, that some of his conclusions should be rejected by mankind generally; or even that the vulgar should keep aloof from his system as a whole. (p. 490)

These consequences, which Williams aptly says Sidgwick "pursued with masochistic thoroughness" (1985, p. 108), are the upshot of Sidgwick's pioneering discussion of utilitarianism's violation of the publicity condition.

[3] Unfortunately it appears that Sidgwick may misstate, slightly, the utilitarian argument for maintaining the common opinion that an act which would be bad if done openly is not rendered good by secrecy. For if the agent did not hold this opinion, then as long as the act were secret, he would not think he had done anything wrong, and his natural disposition to hide his wrongdoings would not be supported—and what Sidgwick says we should take care to avoid (supporting this natural disposition) would not be at risk of occurring. But although Sidgwick's invoking of this natural disposition fails to establish the connection between the common opinion and the operation of the "social disapprobation" sanction, the connection can easily be made in any of several other obvious ways. For example, agents who believe that secrecy would ensure the rightness of certain acts would be likely to keep even wrong acts secret (thinking that those acts are right), thereby interfering with the operation of the sanction of social disapprobation.

2.3 Just as the issue with which Sidgwick wrestles is that of utilitarianism's being esoteric (not, on the one hand, being self-effacing or, on the other hand, being in violation of some stronger requirement such as Rawls's publicity condition), so prominent critics of utilitarianism tend to focus on this intermediate of the three versions of the publicity condition. Baier, for example, writes that

> An esoteric code, a set of precepts known only to the initiated and perhaps jealously concealed from outsiders, can at best be a religion, not a morality. . . . 'Esoteric morality' is a contradiction in terms.[4] (1965, p. 101)

Similarly, Hodgson writes that if certain assumptions are granted, "it would mean simply that universal and correct application of act-utilitarianism could not persist, because it would involve rejection of act-utilitarianism by at least some persons" (p. 46); and he implies that this result would discredit utilitarianism. Somewhat more bluntly, Williams affirms the "capacity for utilitarianism . . . to annihilate itself" on the basis of the fact that "if utilitarianism is true, and some fairly plausible empirical assumptions are also true, then it is better that people should not believe in utilitarianism" (1972, pp. 95–98). Rescher, finally, writes that "it would surely put the utilitarian in an untenable position to concede that his moral theory is not self-sustaining, that it enjoins him to teach and foster a moral theory at variance with itself" (1975, p. 79).

Despite these criticisms, several contemporary defenders of consequentialist theories have willingly declared their theories' esoteric character (typically in order to deflect charges of pragmatic ineffectiveness). Hare, for example, writes that his moral theory may recommend that a police officer make it "a matter of principle never even to contemplate" certain practices such as torture, even though there may be special cases in which engaging in torture would bring about the best consequences (1963, p. 44).[5] Similarly, Parfit argues that egoism would instruct an agent to adopt whatever motives, dispositions, and beliefs—including beliefs about rationality—

[4] Baier goes on to say that this position (that " 'esoteric morality' is a contradiction in terms") yields three other criteria of moral rules" (1965, p. 101), one of which is that moral rules must not be "self-defeating" (1965, p. 102). As Baier uses the term, "A principle is self-defeating if its point is defeated as soon as a person lets it be known that he has adopted" (1965, p. 102). This notion of self-defeat is, clearly, stricter than the simple one of pragmatic ineffectiveness operative in our two previous chapters.

[5] He also writes, "when act-utilitarian reasoning is employed by ordinary humans it may be expected to lead to trouble. They would do better to be guided by simple general principles or by their consciences for most of the time" (1973, p. 60). And again: "there are excellent act-utilitarian reasons for an educator to bring up his charges to think intuitively on most occasions on the basis of a high-quality set of principles selected by critical thinking. This applies equally to self-education" (1976, p. 227). See also Hare (1979, pp. 232–233 and pp. 239–242).

the having of which by the agent in his circumstances would benefit him the most (1984, pp. 9–10); and he makes similar claims about act consequentialism (1984, pp. 40–41).[6]

To be sure, not all defenders of consequentialist theories concede their esoteric character.[7] Smart, for example, writes that

> the great danger to humanity comes nowadays on the plane of public morality—not private morality. There is a greater danger to humanity from the hydrogen bomb than from an increase of the divorce rate, regrettable though that might be, and there seems no doubt that extreme utilitarianism makes for good sense in international relations. . . . I myself have no hesitation in saying that on extreme utilitarian principles we ought to propagate extreme utilitarianism as widely as possible. (p. 348)[8]

And Smart is not alone.[9] But whether this position is defensible or not, my principal concern is this chapter is not whether straightforwardly maximizing theories such as egoism and act consequentialism violate the publicity condition in the way that their critics allege, but whether, assuming that they do, they ought then to be regarded as inferior to their indirectly maximizing rivals.

3 The case for the publicity condition

3.1 As we saw in the last section, the publicity condition is frequently mentioned as a requirement that it may or may not be reasonable to impose on normative theories.[10] As a result,

[6] He mentions, though, that "it is unlikely that [act consequentialism] is wholly self-effacing. It would at most be partly self-effacing. . . . It might make the outcome better if some people did not believe [act consequentialism]; but it is unlikely that it would make the outcome better if [act consequentialism] was believed by no one" (1984, p. 41). So, Parfit, too, seems to regard the possibility of a theory's being esoteric, not outright self-effacing, as the most pressing issue.

[7] Nor do all critics of them relentlessly insist on their esoteric character. It is an interesting irony that Rawls, even while construing the publicity condition in the strongest of the three forms earlier distinguished, seems to imply that utilitarianism may satisfy it. Referring to the set of constraints containing the publicity condition, he writes, "I assume that they are satisfied by the traditional conceptions of justice" (1999b, p. 113). Presumably he counts utilitarianism among these, since he refers to it as "[d]uring much of modern moral philosophy the predominant systematic theory" (1999b, p. xvii) and includes it on the list of the alternatives among which the parties in the original position have to choose (1999b, p. 107).

[8] He concedes, though, that "Sidgwick had respectable reasons for suspecting the opposite" (p. 348).

[9] Langenfus, for example, maintains that consequentialism is not completely self-effacing (pp. 481–482). Brink writes that "In the actual world utilitarianism satisfies the publicity constraint" (p. 428), but in claiming this he seems to assume that an agent can regard utilitarianism as binding while being committed to non-utilitarian moral rules in such a way that they are, for the agent, something more than mere rules of thumb. Because this latter commitment may amount to utilitarianism's no longer being the agent's decision procedure, such an agent would run afoul of the publicity condition as understood here.

[10] Indeed, as Hooker writes, "The literature on the 'publicity condition' is voluminous" (2000, p. 85, n. 14; see also 1990, p. 72, n. 20). Hooker also provides a generous list of references.

it has been the subject of considerable debate, and many attempted defenses of it have been given. In this section, we shall critically examine some of these defenses. In doing so, we shall, where possible, construe the defense under consideration as supporting the weakest (and, hence, most plausible) of the versions of the publicity condition distinguished in the previous section: the ban on self-effacing theories.

3.2 One defense of the publicity condition is suggested by the claim of Baier's quoted above: " 'Esoteric morality' is a contradiction in terms." Even more plausible (since weaker) is the claim that a *self-effacing* morality is a conceptual impossibility, and it might be thought that a self-effacing theory of rationality is equally inconceivable—however conceivable self-effacing *pseudo*-theories may be. But as Scheffler explains in discussing the bearing of the publicity condition on utilitarianism and other forms of consequentialism, even if such claims could be substantiated, they would not endow the publicity condition with the force and scope that their defenders intend for it to have:

> If 'morality' is defined in such a way as to include the publicity condition, and if a thoroughgoing consequentialism dispenses with the publicity condition, then talk about the relative merits of consequentialist and non-consequentialist moral principles can simply be recast as talk about the relative merits of consequentialist principles on the one hand and moral principles on the other. By itself, no simple appeal to meaning is capable of showing that there is something wrong with consequentialism's apparent willingness to violate the publicity condition. (p. 47)

Indeed it seems that an appeal to meaning is bound to be unconvincing except to someone already convinced of the propriety of the condition in question.

3.3 It is notable that Rawls, who (as we saw) defends the publicity condition even in the strongest of the three versions distinguished in the last section, declines to deploy a definitional argument. Instead, he writes that

> There are certain formal conditions that it seems reasonable to impose on the conceptions of justice that are to be allowed on the list presented to the parties. . . . I do not claim that these conditions follow from the concept of right, much less from the meaning of morality. (1999b, p. 112)

And one might think that "it seems reasonable" to impose similar conditions on theories of rationality. Then, as if to emphasize the difference between his approach and Baier's, Rawls adds that "by itself, a definition cannot settle any fundamental question" (1999b, p. 113).

Rawls's approach, being more modest than Baier's, is not vulnerable to precisely the same reply as Baier's. But in its modesty it is vulnerable to an even simpler reply. As Scheffler writes,

> once it is said that the condition is just something it 'seems reasonable' to expect an acceptable moral conception to satisfy, and that the adequacy of the condition must ultimately be assessed in the light of the moral conception it leads us to, the consequentialist can simply deny that the condition 'seems reasonable' to him. (p. 47)

And something parallel may be said on behalf of theories of rationality. So although Rawls consciously avoids the lure of Baier's definitional argument, he does not manage to replace it with anything stronger.[11]

3.4 A third approach to defending the publicity condition begins with the following observation: a normative theory that violates the publicity condition is one that requires agents to cultivate and to maintain beliefs that the theory itself implies are false—beliefs, for example, about what is the best theory of morality or rationality, or (what will be implied by such beliefs) about what acts are right and wrong or rational and irrational.[12] This observation, when conjoined to the thought that any normative theory that can be so described must *ipso facto* be unacceptable, implies the unacceptability of any normative theory that violates the publicity condition.

To be sure, this thought does have some intuitive appeal. For it is natural to think, especially from a philosophical point of view, that the only beliefs that we can have good reasons for cultivating and maintaining are *true* ones. And so while we are accustomed in philosophy to encountering theories (normative and otherwise) that have implications that *we* regard as false, we may react with particular suspicion to a theory that recommends beliefs that *it* implies are false. Such a theory may seem not only mistaken in some way, but also—and more seriously—guilty of some sort of philosophical bad faith.

But this assessment should look less appealing in the light of the following fact: such a theory (one that recommends beliefs that it implies are false) needn't be guilty of the logical sin

[11] Given the weight that Rawls puts on the notion of the *reasonable* in his later work—especially in his lectures on Kantian constructivism (cited above, in note 2)—it might be thought that Rawls's earlier assertion of the reasonableness of the publicity condition is more pregnant with meaning than I acknowledge. But in my view, Rawls's later remarks on the notion of the reasonable fail to provide the ingredients for a fuller or further argument for the publicity condition.

[12] That a person's acceptance of a theory involves not only her values and motivations, but also her beliefs, is emphasized by Langenfus (p. 479).

of implying that those beliefs are true (something that would be a logical sin because we have already supposed that the theory in question implies that those beliefs are *false*). Rather, the theory may recommend those beliefs for reasons that do not presuppose that those beliefs are true. For example, the theory may recommend those beliefs not for epistemic reasons, but for pragmatic, or practical reasons.[13]

Such reasons are, obviously, what self-effacing theories appeal to. Parfit states this point concisely in his observation that egoism "is a theory about practical not theoretical rationality" (1984, p. 23). In other words, it is a theory about advancing one's aims, not about correcting one's beliefs (except, of course, insofar as having correct beliefs contributes to the advancement of one's aims). As a result, egoism "may tell us to make ourselves have false beliefs" (1984, p. 23). Parfit makes the same point in another way in regard to act consequentialism:

> there are two questions. It is one question whether some theory is the one that we *ought morally* to try to believe. It is another question whether this is the theory that we *ought intellectually* or *in truth-seeking terms* to believe—whether this theory is the true or best justified theory. (p. 43)

Once these two questions are distinguished, then a theory that recommends beliefs that it implies are false may continue to seem (as Sidgwick said) paradoxical; but it cannot be dismissed as unacceptable on logical grounds.

Distinguishing these two questions also exposes the fallacy implicit in a superficially clever, but ultimately misleading, apparent dilemma for defenders of self-effacing theories. Here is how Williams constructs the alleged dilemma:

> [I]f utilitarianism is true, and some fairly plausible empirical assumptions are also true, then it is better that people should not believe in utilitarianism. If, on the other hand, it is false, then it is certainly better that people should not believe in it. So, either way, it is better that people should not believe in it. (1972, p. 98)

Clearly, the 'better' in the first premise—the one that alludes to utilitarianism's self-effacement—is a pragmatic 'better', not an epistemic one. And the reverse is true of the 'better' in the second premise—the one predicated on utilitarianism's falsity. Thus, no matter which way

[13] In pointing out that a theory of the kind under discussion may avoid logical inconsistency by appealing to pragmatic rather than epistemic reasons, I do not mean to imply that there cannot also be *epistemic* reasons for holding false beliefs. For example, there may be epistemic reasons for holding those beliefs that best enable one to acquire true beliefs—even if some or all of those enabling beliefs are themselves false. For further discussion, see Heil (1983, pp. 754–757; and 1992, pp. 47–48).

the 'better' in the conclusion is disambiguated, the conclusion will lose the support of at least one of its two premises, and will not remain standing.[14]

3.5 A fourth argument for the publicity condition turns on the thought that any theory that violates the publicity condition is objectionable because it requires agents to dispose themselves to act in ways that the theory itself proscribes (as irrational, immoral, or in some other way, depending on the character of the theory).[15] Now this claim has obvious parallels to the last one—that a theory ought not to require agents to cultivate beliefs that the theory itself implies are false—but cannot be dismissed as easily. For we dismissed the last claim by pointing out that it is essentially an epistemic concern, and that a normative theory is concerned with practice. But the present claim, by referring to proscribed acts instead of false beliefs, strikes closer to home.

But still it may be held off. For when a theory requires an agent to dispose himself to perform certain acts, then the theory needn't imply that those acts are rational, right, or otherwise acceptable. So, in requiring an agent to dispose himself to act in ways that the theory itself proscribes, the theory needn't commit the obvious logical sin of implying that those acts are right, or that there are any other practical reasons for performing those acts. All the theory is committed to is that there are practical reasons for being *disposed* to perform those acts. In terms introduced in chapter II (in subsection 6.3 of that chapter), those reasons may have to do with the autonomous effects, rather than with the direct effects, of being so disposed.

Still there may seem to be something odd about a theory that can be tolerant, in this way, of proscribed conduct. But note that, for standard versions of straightforwardly maximizing theories, acting rationally or morally is only a means to whatever end the theory is organized around. Of course, it is possible to construct variants of these standard versions in which rational or moral action itself is part of the end or one of the ends (though a problematic regress seems to lurk in the thought that it could be all of the end or the sole end), but such variants are not our concern here. Proscribed acts are not worth avoiding except insofar as they detract from the

[14] Admittedly, the premises do support the following inference: that there is *some* sense of 'better' in which it is better that people not believe in utilitarianism. But utilitarianism itself already implies this, insofar as it admits to being self-effacing. It should also be admitted that Williams's remarks may well be enthememetic for an argument that cannot be dismissed so easily. For another reaction to Williams's blurring of the pragmatic and the epistemic, see Scheffler (p. 51).

[15] That egoism "might tell us to cause ourselves to be disposed to act in ways that egoism claims to be irrational" has been observed by Parfit (1984, p. 13), and Kavka remarks that act consequentialism may require an agent to "corrupt himself" (1978, p. 295).

advancement of the aims of the theory. And of course it may happen that the best way for an agent to advance certain aims is for the agent to embark on a trajectory that involves many more proscribed acts than some other trajectory available to the agent. For example, it may happen that the best way for an agent to advance her aims is to embark on a non-egoistic trajectory that involves many more self-sacrificing acts, which egoism would deem irrational, than some other trajectory available to her. A theory could disallow such a trajectory for that reason only if, as just mentioned, it had some special concern with irrational or immoral action as an end, and not just as a means. But if it regards action only as a means, then although irrational or immoral action would not be worth doing in and of itself, it would not be worth avoiding to *such* an extent that the trajectory of which it is a part ought to be forsworn.

3.6 A fifth, and final, argument for the publicity condition is that theories that violate it also violate the dictum that 'ought' implies 'can'. To see how this might be thought to be the case, note that any normative theory will imply that an agent ought never to perform any act it proscribes. A theory of rationality will imply that an agent ought never to perform any act that it implies is irrational, and a theory of morality will imply that an agent ought never to perform any act that it implies is immoral. And yet if a theory is self-effacing, then it seems to imply that some agent must, at some time or another, perform an act it proscribes. For if the theory is self-effacing, then it requires some agent to dispose himself to perform acts it proscribes, and this gives rise to a dilemma. If the agent complies with that requirement, then presumably the agent will go on to perform some proscribed acts; and if the agent violates that requirement, then the agent performs a proscribed act right then. Either way, the agent performs at least one proscribed act. But the theory says that the agent ought never to do that. So the agent simply *cannot* do what, according to the theory, the agent *ought* to do. Therefore, a theory that violates the publicity condition violates the dictum that 'ought' implies 'can'.[16]

Now this is a difficult point, and one to which certain specious replies may seem adequate. One such reply is as follows. The fact that an agent disposes herself to perform some proscribed acts does not mean that she cannot avoid performing proscribed acts, for she may be able to secure the benefits of being so disposed, and then cease to be so disposed, before the time for performance arrives. For example, I may be disposed to grant patents, even when doing so

[16] A similar statement of this argument is provided by Parfit (1984, pp. 13–16 and pp. 35–37). For a discussion of Parfit's account of it, see Mintoff.

would be a net loss—thereby eliciting valuable inventions—and then cease to be so disposed, before the time for performance arrives—thereby bringing about the best available outcome. Now why is this reply inadequate? Because although such a turn of events is possible, agents do not have such control over their mental states that it can be said of them that they *can* do that in any non-trivial sense of ‘can’, any more than bowling a 300 game is something that I can do in any non-trivial sense of ‘can’, even though it is perfectly possible for that to happen, too. (I do occasionally bowl a strike.) The dictum is that ‘ought’ implies ‘can’, not that ‘ought’ implies ‘might possibly happen’.

So this reply gets derailed by its reliance on an objectionably trivial construal of ‘can’. But it is on the right track insofar as it suggests that the notion of ‘can’ needs to be qualified in some way. A better reply proceeds as follows. The fact that an agent disposes herself to perform some proscribed acts does not mean that she cannot avoid performing proscribed acts in the sense of ‘can’ at work here. For even when she is disposed by her decision procedure to do something, she can still do something else in the sense of ‘can’ at work here. For one principle that regulates our use of ‘can’ in this context is that what an agent can do in a particular situation is not dependent on what her decision procedure disposes her to do. Indeed this principle is insisted upon even by the critics of straightforwardly maximizing theories (who presumably would also welcome a showing that such theories violate the dictum that ‘ought’ implies ‘can’)—since if it is denied, then it must be denied that an egoist never acts irrationally and an act consequentialist never acts wrongly, since there is nothing that they *can* do other than what their decision procedures dispose them to do. So the fact that an agent’s decision procedure disposes her to do what she ought not to do is compatible with the claim that she can do what she ought to do.

Now it may be thought that I am appealing to a watered-down notion of an agent’s being disposed in a certain way, so that the agent is understood to be capable of resisting her disposition, overruling her decision procedure, and acting for other reasons. And such an appeal would obviously be objectionable, since it would rob the notion of a disposition of the force that it needs to have if an agent’s disposition is supposed to make a difference to the opportunities she faces and the acts she performs. But my reply depends on no such appeal. For suppose that we say that an agent’s disposition, or decision procedure, is absolutely determinative of her conduct. It is still the case that in the sense of ‘can’ at work here, she can do otherwise just as much as if her disposition exercised a much weaker influence on her. For in order to ascertain whether an agent can do something or not, the appropriate question is not, “Given the agent’s decision procedure, might she decide to do this?” For, as I said in the last paragraph, if that were the question, then no agent with a fully determinative decision procedure can ever do anything other than what he does do, and no agent with a fully determinative decision procedure can ever act irrationally or immorally. Rather, in order to ascertain whether an agent can do something, the appropriate question is, “Abstracting from the decision procedure that will determine her choice, is this among her options? Is this an act that should be surveyed by her decision procedure before it selects something as what is to be done?”[17]

So even when an agent is disposed by her decision procedure to perform some act, she can still perform some *other* act in the relevant sense of ‘can’. Admittedly, there is a sense of ‘can’ in which it is *not* the case that she can do something else, but that sense of ‘can’ does not concern us. So, the fact that an agent disposes herself to perform some proscribed acts does not mean that she cannot avoid performing proscribed acts. On the contrary, she can; indeed she can do what the theory says she ought to do. So a theory needn’t violate the dictum that ‘ought’ implies ‘can’ if it violates the publicity condition, and the publicity condition is not implied by that venerated maxim.

3.7 In this section, I have outlined and refuted five lines of defense for the publicity condition: that ‘esoteric morality’ is a contradiction in terms (subsection 3.2), that the publicity condition is a reasonable one to impose on normative theories (subsection 3.3), that a theory is objectionable if it implies that agents ought to cultivate beliefs that the theory implies are false (subsection 3.4), that a theory is objectionable if it implies that agents ought to dispose themselves to perform acts that the theory proscribes (subsection 3.5), and that the publicity condition is implied by the dictum that ‘ought’ implies ‘can’ (subsection 3.6). Several promising putative justifications for the publicity condition turn out, then, to leave it without solid support.

[17] Compare Quinton’s suggestion that “what an agent *can* do, in the sense required for it to be the case that he ought to do it, is what he could be induced to do by sanctions” (p. 7).

4 Two refutations of the publicity condition

4.1 The publicity condition is not only, as I argued in the last section, in need of solid support. It is also vulnerable to direct attack, as a demonstrably unreasonable requirement to impose on normative theories. In this section, I offer two separate arguments, each of which I contend is sufficient to refute the publicity condition even in its weakest, and hence most plausible, version: the ban on self-effacing theories.

4.2 First, the publicity condition is simply question-begging against any theory that violates it. To see this, recall what it means for a theory to be self-effacing: a theory is self-effacing if it requires every agent in the group to which it applies not to subscribe to it. But how, exactly, would a theory require this? Note that subscribing or not subscribing to a theory is not like stealing or not stealing an unattended bicycle. One can (normally) choose, at will, to steal or not to steal; but one cannot just choose, at will, to subscribe or not to subscribe to a particular theory, any more than one can just choose, at will, to subscribe or not to subscribe to the belief that the earth is flat. So, for a theory to require every agent in the group not to subscribe to it—which, as recalled above, is what's involved in a theory's being self-effacing—must not mean for it to require every agent in the group simply to refrain from subscribing to it, as it might require every agent in that group to simply refrain from stealing bicycles.

What, then, *does* it mean for a theory to require every agent in the group to which it applies not to subscribe to it? If subscribing or not subscribing to a particular theory is not something that an agent can simply choose to do, then what can it mean for a theory to require such subscription or non-subscription? The key to answering this question lies in seeing that although there may be no way in which agents can *directly* control their states of subscription and non-subscription (as we might call them), there are familiar ways in which they can *indirectly* control them: they can set up social and educational institutions in certain ways, they can decide to subject themselves to certain influences rather than others, and so on. It is also important to see that many of these means of control are going to be ones that a person can exercise over others as well as over himself.

With these observations in place, we are in a position to say what it means for a theory to require every agent in the group to which it applies not to subscribe to it. It is for the theory to require some agent or agents to perform some act or acts that would cause (whether intentionally or not) every agent in the group not to subscribe to it, or—and this is a mouthful, but it's another way of saying the same thing—by forbidding some agent or agents to perform the only act(s) that would cause it to continue to be the case that some agent in the group continues to subscribe to it. For example, an agent may be situated such that one of the acts open to her would result in every agent's not subscribing to some theory, and that theory may happen to select *that* act as the one that she ought to perform. Or an agent may be situated such that only some of the acts open to him would save some theory from being effaced, and that theory may happen to forbid him to perform any of those. For our purposes the important aspect of all this is that if and when a theory violates the publicity condition, it does so in virtue of the *content of its prescriptions*. The publicity condition, then, amounts to a substantive constraint on the prescriptions that a theory may issue.

One implication of this understanding of the publicity condition is to call into question the standard interpretation of it as an innocuous *formal* constraint on normative theories.[18] For its evident sensitivity to the content of moral theories rather suggests that it is a *material* one. Now one might argue, on the contrary, that this constraint is really just a formal one (and so not a material one), since it is relativized to each theory (saying that each theory must not imply that agents ought not to subscribe to *it*, where the 'it' varies from theory to theory). But I shall not linger over this dispute. For I am more concerned with another implication of the understanding of the publicity condition just adumbrated: that the publicity condition, by discriminating among theories on the basis of the content of their prescriptions, begs the question of what the correct prescriptions of morality or rationality are. As Railton writes,

> any such condition would be question-begging against consequentialist theories, since it would require that one class of actions—acts of adopting or promulgating an ethical theory—*not* be assessed in terms of their consequences. (p. 155)

Similarly, Brink writes that

> the publicity condition simply begs the question against teleological theories. Whether the true moral theory should be recognized, taught, or recommended as a decision procedure are themselves practical questions the answers to which, the teleologist claims, depend upon the intrinsic and extrinsic value that this sort of publicity produces. (p. 428)

[18] "The Formal Constraints of the Concept of Right" is the title of the section in *A Theory of Justice* in which Rawls formulates his publicity condition (1999b, p. 112).

Brink's conclusion sums up the essential point: "The publicity constraint, therefore, must be construed as a substantive moral claim" (p. 428).

The point is not, of course, that a substantive normative claim cannot bear on the evaluation of a normative theory. On the contrary, such a claim may be eminently relevant to the evaluation of a normative theory. But any such claim offered as dispositive of a normative theory, consequentialist or otherwise, ought to be embedded in a competing normative theory—or at least a sketch of one. For if such a claim is asserted as an independent and incontrovertible normative truth—as the publicity condition is when it is asserted as a freestanding requirement that it is reasonable to impose on normative theories—then it is bound to seem, and to be, blatantly question-begging.

4.3 My second refutation of the publicity condition is somewhat more involved, but culminates in an equally simple claim: that this requirement is unreasonably demanding. The argument proceeds by showing that the range of normative theories that violate the publicity condition is wider—indeed is nearly exhaustive of the range of available normative theories—than one might have initially expected. This result, when coupled with the thought that only an unreasonably demanding requirement rejects so wide a range of the available normative theories, underwrites the conclusion that the publicity condition is an unreasonably demanding requirement to impose on normative theories.

First, consider alternatives to egoism such as constrained maximization and resolute choice. To be sure, these theories may violate the publicity condition to a lesser extent than egoism, in the sense that many of the circumstances in which egoism would be self-effacing would not necessarily be circumstances in which they would be self-effacing. For example, in circumstances in which agents stand to gain from being able to make promises and threats, or to form other intentions to perform non-optimal acts (such as drinking toxins), constrained maximization and resolute choice would not follow egoism in being self-effacing, since they are compatible with the agent's making promises and threats in a way that egoism is not. But there may still arise cases in which even these theories are self-effacing. For example, in the case of the tyrant (introduced in subsection 3.1 of the last chapter), constrained maximization and resolute choice would presumably direct an agent to adopt a tyrant-placating decision procedure. So these rivals eventually join egoism in violating the publicity condition.

What about rule-based alternatives to straightforwardly maximizing theories, such as rule egoism and rule consequentialism? Let us focus on rule consequentialism, since the single-agent case of rule-egoism follows trivially once the multiple-agent case of rule consequentialism is worked out. To see how rule consequentialism violates the publicity condition, reflect on how act consequentialism violates it. The agents in some act-consequentialist society find themselves in circumstances in which better results would come from their all not subscribing to act consequentialism. (This may happen for any of a number of reasons; some examples, including monetary policy and military deterrence, were discussed in chapter II.) What makes act consequentialism self-effacing in such circumstances is that they are circumstances in which what's optimal is for the agents to transform themselves into agents who do not regard morality as requiring them to perform optimal acts.

Similar circumstances may arise for rule consequentialism—the theory that says that an agent ought to perform whatever act is prescribed by those rules whose general acceptance would have consequences that are better than the general acceptance of any other rules. For there is no reason, in principle, why the agents in some society cannot be situated so that one of their optimal rules prescribes that, in the circumstances in which they now find themselves, a certain agent perform an act that results in all of the agents' being caused not to subscribe to rule consequentialism. To see this, note that it is not necessary, in order for rule consequentialism to be self-effacing, for any of its rules to explicitly prescribe that the theory be effaced in certain circumstances; all we need to suppose is that the agent finds himself in a situation in which he has only two options, with one option being in violation of his society's optimal rules and the other option having, as one of its consequences (either as an intended result or as a side-effect), the effacement of the theory. In short, all that is needed in order for rule consequentialism to violate the publicity condition is for it to be possible for agents to happen to find themselves in circumstances in which their optimal rules require conduct that happens to result in the theory's effacement. And this is obviously possible.

Now we saw in the last chapter that advocates of rule-based theories of morality and rationality may object to the pragmatic ineffectiveness of act-based theories, and that defenders of act-based theories may reply not only by questioning the validity of pragmatic-effectiveness criterion but also by pointing out that rule-based theories are also pragmatically ineffective, due to their extreme abstractness. Now we have seen that the same sort of exchange is possible in regard to the publicity condition: if advocates of rule-based theories criticize act-based theories for violating this condition, then defenders of act-based theories may reply not only by questioning the validity of this condition, but also by pointing out that rule-based theories also violate it.

4.4 At this point it has been shown only that maximizing theories—whether straightforward or indirect—violate the publicity condition. And a defender of the publicity condition may just say, “So much the worse for maximizing theories, of whatever stripe, if the publicity condition implies their rejection. There are other theories out there.” But there is more to be said about the objectionably strong implications of the publicity condition.

Consider now a broader class of theories: those that, while not necessarily *maximizing* in any way, are just responsive enough to consequences to include a disaster-avoidance provision requiring agents to avert disasters if they are able to do so without imposing very great costs on themselves, where a disaster is understood (as in subsection 5.3 of the last chapter) to be an outcome that is much worse than some alternative.[19] Note that this class of theories, while containing most standard forms of consequentialism, is also broad enough to contain many theories that would not be called consequentialist, since the responsiveness to consequences that characterizes the theories in this class is so weak. A recognition of just *how* weakly consequentialist the disaster-avoidance provision is emerges from a consideration of two features of it. First, not only does the provision not require agents to bring about the best possible outcomes; it does not even require them to avoid the worst possible outcomes, except in those cases in which the worst possible outcome is *much* worse than some alternative. Second, it excuses agents from this requirement whenever fulfilling it would require them to shoulder heavy burdens. So the disaster-avoidance provision is a very weakly consequentialist principle—weak enough, in fact, to certainly be included in such non-consequentialist theories as common-sense morality. (Indeed common-sense morality may include an even more consequentialist disaster-avoidance principle than the one I have sketched. But I focus on the one stated in order to discuss as large a class of theories as possible.)

The class of disaster-avoiding theories, then, is a broad one, including non-consequentialist theories as well as consequentialist ones. But, remarkably, *every* theory in this

[19] A class of theories similar to this one, but not as quite broad, is defined by Kavka for a similar purpose (1978, p. 287).

class violates the publicity condition. To see this, let T be some theory in this class. Now suppose that an agent finds himself in a situation in which he has only two options, with one option being not only worse than the other according to T, but also *enough* worse than the other for T to count it as a disaster. Since T is (*ex hypothesi*) a disaster-avoiding theory, T requires the agent to choose the second option (the disaster-avoiding one). But suppose also that the second option involves—either as a means to its intended result or as a side-effect—causing everyone in the group to which T applies not to subscribe to T. Then T, by requiring the agent to choose the second option, violates the publicity condition. Thus every disaster-avoiding theory violates the publicity condition.[20]

4.5 Already it is clear that the publicity condition is more demanding than it might have initially seemed, since it rejects non-consequentialist theories such as common-sense morality as well as consequentialist ones. But one might still think that it is not *unreasonably* demanding, on the ground that it can be satisfied by certain normative theories that abjure consequentialist considerations altogether. One might think, for example, that Kant's moral theory satisfies the publicity condition, not only because of its rigorously non-consequentialist character, but also because Kant is explicitly credited with having developed a moral theory in the spirit of the publicity condition. Rawls, for example, writes that "The publicity condition is clearly implicit in Kant's doctrine of the categorical imperative insofar as it requires us to act in accordance with principles that one would be willing as a rational being to enact as a law for a kingdom of ends" (1999b, p. 115). So it would be telling indeed if Kant's theory could be shown to violate the publicity condition.

But Kant's moral theory can be shown to do just this, by way of an argument analogous to the one offered in reference to disaster-avoiding theories. Begin by supposing that some agent finds herself in a situation in which she has only two options, with one option being a textbook

[20] Although it must be admitted that a case with the structure just described is unlikely to arise in practice, the bare logical possibility of one is entirely sufficient for our purposes. And so I shall not tax the reader's patience by indulging in the elaborate stage-setting that developing an example of even modest plausibility would require. But those readers who wish to consider *some* example, however fanciful, may entertain the following scenario. An asteroid is headed towards Earth, and the one scientist who knows how to stop it refuses to do so unless everyone on the planet repudiates whatever moral theory he or she currently subscribes to and subscribes, instead, to the scientist's own creed (whatever that happens to be). Now there are some people who subscribe to one particular disaster-avoiding theory, and it is in the power of one of them to cause herself and the rest of them to repudiate it in favor of the scientist's creed. Because she can do this without substantial cost to herself, and because her inaction would result in an outcome that would be *much worse* (according to her theory) than her acting so as to reshape the moral commitments of herself and others, her theory requires her to act in that way. Thus, her theory requires its own effacement—which is to say, of course, that it violates the publicity condition.

example of an act that violates the categorical imperative, such as lying. But suppose also that the second option involves (again, either as a means to its intended result or as a side-effect) causing all rational beings not to subscribe to Kant's moral theory.[21] Then, ironically, not even Kant's moral theory—which is thought to be the natural home of the publicity condition—turns out to satisfy it.[22]

4.6 It might appear that the conclusion towards which we are driving is that *no* normative theory satisfies the publicity condition. Indeed this claim seems to be the upshot of Brink's statement that "For any moral theory, there are possible circumstances in which its recognition and application would satisfy the theory worse than recognition and application of some alternative theory" (p. 429). But we cannot endorse a claim quite this strong. For there is one class of normative theories that satisfy the publicity condition, which as far as I know has not been previously identified as such—though before specifying it I should mention that it is a rather trivial one (and one that, therefore, Brink and others may quite reasonably have regarded as not worth recognizing). This class consists of those normative theories that specifically and absolutely forbid agents to act in a way that causes their effacement. Such theories satisfy the publicity condition either by requiring *only* that agents not act in a way that causes their effacement, or by imposing other duties on agents but making those other duties lexically subordinate to the duty of non-effacement.

But no *other* theories than these satisfy the publicity condition. To see this, consider one last argument of the form already used in reference to disaster-avoiding theories and Kant's moral theory. Let T be some theory that not only imposes on agents some duty other than that of non-effacement, but also neglects to make this other duty lexically subordinate to that of non-effacement (either by requiring non-effacement but neglecting to give it lexical priority, or by neglecting to require non-effacement at all). This means that there are circumstances in which this other duty outweighs, trumps, or otherwise takes precedence over that of non-effacement. Now suppose that an agent finds himself in such circumstances and, moreover, in a situation in

[21] As before (see footnote 20), although a case of this kind is unlikely to arise in practice, the bare possibility of one is sufficient for our purposes. And having sketched the asteroid example above, I shall refrain from indulging in further flights of fanciful scenario-building. For another account of how Kant's moral theory violates the publicity condition—an account with a rather different argumentative strategy—see Brink (p. 429, n. 27).

[22] A further irony is that Kant himself alerts us to the possibility that the publicity condition may have surprisingly strong substantive implications, in his derivation from it of the claim that rebellion is always wrong (p. 348).

which he has only two options, with one option being in violation of this other duty (say, the duty not to kill innocents). Then T requires the agent to choose the second option (e.g., not to kill any innocents). But suppose also that the second option involves causing everyone in the group to which T applies not to subscribe to T. Then T, by requiring the agent to choose the second option, violates the publicity condition. Thus, any theory that does not give lexical priority to non-effacement violates the publicity condition.

4.7 Let us now sum up the results of the last few subsections. We have seen that the publicity condition rejects not only straightforwardly and indirectly (including rule-based) maximizing forms of consequentialism (subsection 4.3), but also theories with mildly consequentialist components such as a disaster-avoidance provision (subsection 4.4), non-consequentialist theories such as Kantianism (subsection 4.5) and, in fact, any theory that is not specially designed with the publicity condition in mind (subsection 4.6). The sets of theories ruled out by the publicity condition, and these sets' relations to one another, are shown in the following figure.

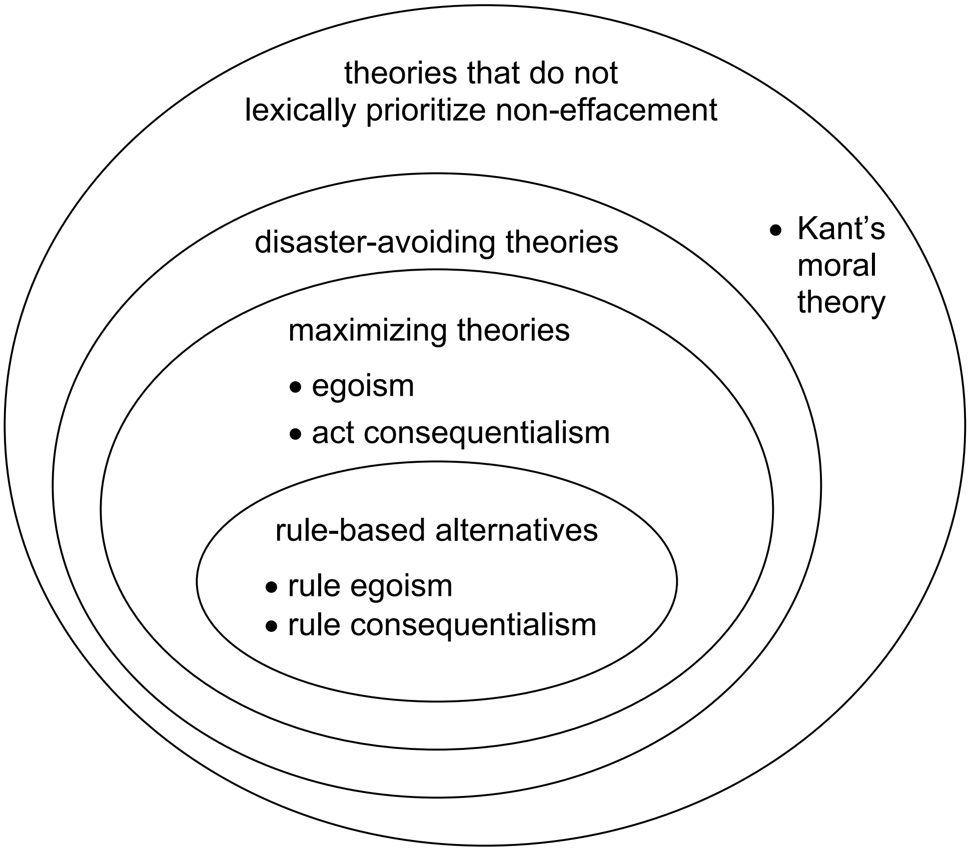

Figure IV.2

Now recall Figure IV.1, which shows how the various versions of the publicity condition are related to each other. Since it, like Figure IV.2, is a diagram of sets of theories, the two diagrams can be combined. The result is shown in Figure IV.3.

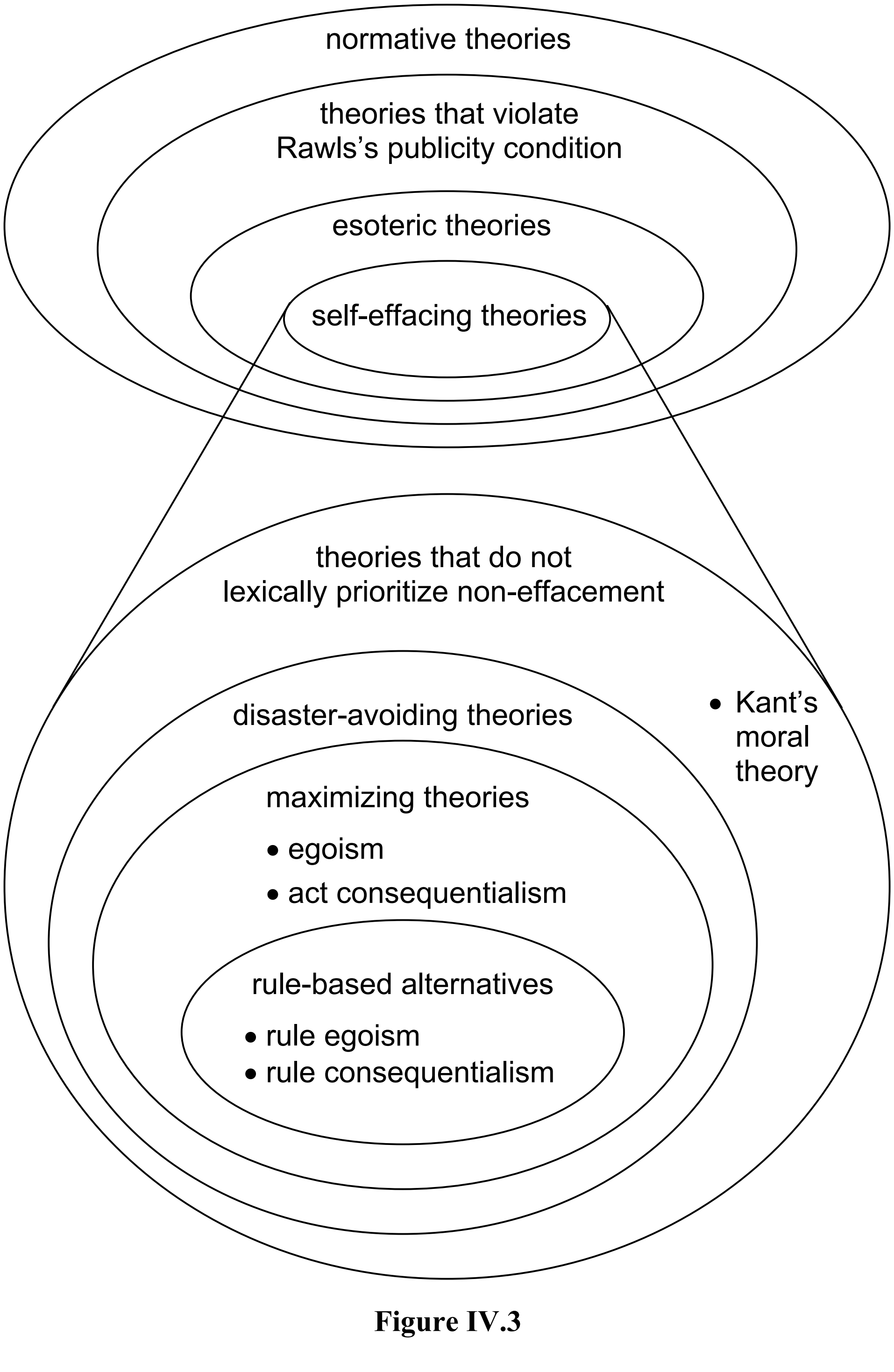

Figure IV.3

What this diagram illustrates, with its diagonal lines connecting the first set of ovals to the second, is that the set of self-effacing theories *just is* the set of theories that do not lexically prioritize non-effacement (which is implied by our finding, above, that all and only theories that lexically prioritize non-effacement satisfy the publicity condition). Since the most plausible version of the publicity condition—the ban on self-effacing theories—rejects all of the theories in the innermost over in the upper part of Figure IV.3, it rejects all of the theories in the lower part of Figure IV.3.

The question that arises now is whether the publicity condition, being as demanding as the foregoing subsections indicate, is a reasonable requirement to impose on moral theories. In essence, the question is this: is it reasonable to insist (as even the most plausible version of the publicity condition does) that a normative theory lexically prioritize non-effacement? To answer this question, let us consider what lexically prioritizing non-effacement entails. It entails consigning all other values—whether standard consequentialist ones such as well-being or traditionally deontological ones such as being truthful and being respectful of others' lives and rights—to lexically subordinate positions. And this, I submit, is an unreasonable demand. Of course I cannot here offer a *conclusive* argument showing the importance of other values relative to non-effacement, but I can invite the reader to turn to any normative theory attracting widespread attention—any in Figure IV.2, at least—for then she will surely find an account of values that rejects the lexical priority of non-effacement (if it grants non-effacement any importance at all). Indeed the granting of lexical priority of non-effacement is such an extreme position that one might go so far as to say that not only does the publicity condition reject many reasonable theories, but also, the only theories it accepts are *unreasonable* ones.

This last point, in addition to showing just how unreasonably demanding the publicity condition is, also allows us to redeem a promissory note issued in the second paragraph of this chapter, to the effect that a non-self-effacing variants of self-effacing theories are certainly possible, but they are bound to be unsatisfactory on other grounds. For a self-effacing theory can be turned into a non-self-effacing one only by completely subordinating *all* of its values and requirements to a non-effacement requirement. Nothing short of that will do. And yet such a radical maneuver would, I have argued, be objectionable on substantive grounds.

We may conclude, then, that because of the obvious unacceptability of the only theories that can be assured of satisfying the publicity condition, the publicity condition is unreasonably demanding. Indeed, since the publicity condition threatens to rule out all but the most publicity-committed normative theories, this condition, far from being an innocuous formal constraint, is a strong and unacceptable substantive constraint—not to mention being a question-begging one (subsection 4.2)—on the values that a theory may embody.

4.8 Before closing this section we should pause to consider an objection to the framework within which the argument of this section proceeds. Recall that the gist of the argument of this section is that the ban on being self-effacing is the weakest, and hence most plausible, of the versions of the publicity condition, and that even it has objectionable implications. Now in reply to this, a defender of the publicity condition might grant that the ban on being self-effacing does have objectionable implications, but might also claim that there is a weaker, and hence more plausible, version of the publicity condition that I failed to articulate in section 2.

What might this weaker version be? Here is one candidate: that a theory is unacceptable not just if it prescribes its own effacement in some possible circumstances, but if it prescribes its own effacement in *actual* circumstances. Then the appeal to hypothetical cases—admittedly so crucial to showing how various are the theories that violate the publicity condition—is blocked, and the publicity condition may not seem so demanding after all. But there are, I believe, stronger reasons that support focusing on what might be called in-principle versions of the publicity condition, such as the ones I set out above, rather than an in-actuality version of the publicity condition such as the one just suggested. First, when other requirements are used to assess moral theories, the versions of them employed are typically in-principle rather than in-actuality versions of them. For example, when the implications of a theory are tested against commonly held intuitions about specific cases (à la reflective equilibrium), the appeal to possible, not just actual, cases, is generally held to be perfectly acceptable. Second—and perhaps underlying the first reason—appealing to an in-actuality version of some requirement, rather than an in-principle version of it, requires the philosopher to answer some complicated questions of psychology and sociology. No very strong thesis about the separation of philosophy from the empirical sciences is needed in order to maintain that such questions are better set aside when significant conclusions can be derived in independence from them.

But the in-actuality version of the ban on being self-effacing is not the only candidate that we have to consider. Here is another: that a theory is unacceptable not just if it prescribes its own effacement in some possible circumstances, but if it prescribes its own effacement in some possible circumstances *without appeal either to incidental factors such as implementation problems or to fanciful expedients such as tyrants*. Underlying this version of the publicity condition is the thought that a theory may be excused for being self-effacing due to certain factors (such as implementation problems and tyrants), but not for being self-effacing due to certain other factors (such as dynamic inconsistency of the sort experienced by straightforwardly maximizing Humean farmers). On this interpretation of the publicity condition, certain theories—such as constrained maximization and resolute choice—may well satisfy that condition, thus leaving us with an interpretation of that condition that may not be too demanding to be reasonable.

But is it reasonable to disregard factors such as implementation problems and tyrants as irrelevant to whether a theory is self-effacing in a way that cannot be excused? Since I have already touched on the relevance of hypothetical cases, I shall not dwell further on the propriety of appealing to tyrant-based cases (though I recognize that the appeal to such cases represents a particularly extreme form of the appeal to hypothetical cases, one that may need its own, further, justification). As for implementation problems, while they are perhaps the least interesting factors from a theoretical point of view, they cannot be dismissed as unimportant. For as I argued in subsection 6.6 of chapter III, the features of human psychology that lead to these problems—lack of information, lack of calculating ability, limited ability to overcome temptation—are inescapable features of the human condition. A very idealized approach to human thought and action would have to be justified before we could accept the proposed weakening of the publicity condition.

5 Conclusion

This chapter has been devoted to an assessment of the publicity condition as a requirement to impose on normative theories. After distinguishing three versions of the publicity condition and documenting the importance accorded to this requirement in the literature (section 2), we saw the inadequacy of several arguments that may seem to justify this requirement (section 3). Finally, we saw that this requirement is both question-begging and unreasonably demanding (section 4).

This is not to say, of course, that the general idea of the publicity condition is one that it would be undesirable for a normative theory to embody. On the contrary, the insistence that it is an unreasonable requirement to impose on normative theories may well be accompanied by the admission that it would be a nice turn of fortune's wheel if the publicity condition were a reasonable requirement. Indeed even given all that has been said above, plenty of room is left for the admission that it is unfortunate that any viable normative theory will, apparently, have to violate the publicity condition. Sidgwick, who (as we saw in subsection 2.2) anticipated so much of the debate over the publicity condition, anticipated even this aspect of it: Parfit writes that "Sidgwick regretted his conclusions, but he did not think regret a ground for doubt" (1984, p. 41); and Parfit expresses much the same sentiment in the epigraph to this chapter. Brink, similarly, writes that

> Publicity is a plausible but revisable substantive moral commitment. A moral theory that violated publicity in the actual world would be less plausible for that reason. But the fact that there are merely possible circumstance in which a moral theory would require violation of publicity is not a fact peculiar to utilitarianism and is not itself, I think, an objection to utilitarianism or to any other moral theory. (p. 429)

Langenfus, finally, puts the point as follows:

> [T]he situation where the vast majority of moral agents would (morally speaking) be precluded from having an explicit or conscious access to the *true* ground of moral obligation would, no doubt, be a disturbing fact. But, however disturbing this might be in terms of "truth-seeking" ideals, on such assumptions, it would continue to be the only *morally* acceptable situation. (p. 488)

So the thesis of this chapter—that the publicity condition is an unreasonable requirement to impose on normative theories—does not imply that the publicity condition is an unreasonable condition for one to *want* theory to satisfy. It only implies that—in this instance, at least—we would be unreasonable to let our wants regarding normative theories control our evaluations of them.

V

Incoherence

> the legislator, or other practitioner, who goes by rules rather than by their reasons, like the old-fashioned German tacticians who were vanquished by Napoleon, or the physician who preferred that his patients should die by the rule rather than recover contrary to it, is rightly judged to be a mere pedant, and the slave of his formulas.
>
> —John Stuart Mill (1843), p. 944 [bk. VI, ch. xii, § 1]

1 Introduction

The last two chapters were devoted to examining the pragmatic-effectiveness criterion and the publicity condition—two criteria that are often thought to establish the superiority of indirectly maximizing normative theories (such as rule egoism, constrained maximization, resolute choice, and rule consequentialism) to their straightforwardly maximizing counterparts (such as egoism and act consequentialism). We saw, though, that these standards do not provide such strong and unequivocal support for indirectly maximizing theories, not only because the standards themselves are of questionable merit, but also because they represent requirements on normative theories that not only straightforwardly maximizing theories, but also indirectly maximizing theories, have trouble satisfying. The last two chapters, then, were devoted to undermining two of the main grounds for preferring indirectly maximizing theories to straightforwardly maximizing theories. My approach in this chapter is different. Instead of identifying and trying to undermine yet another pillar of support that indirectly maximizing theories rely on, I mount an affirmative attack on those theories by characterizing and defending an objection to those theories that, I argue, shows them to be untenable.

I state and explain the objection, which I call the incoherence objection, in section 2. In section 3, I provide an argument for the incoherence objection, devoting sections 4 and 5 to a consideration of remarks made by Gauthier and McClennen that may seem to protect their theories of rationality from the incoherence objection. I conclude, in section 6, with a summary of the chapter.

2 A statement of the incoherence objection

2.1 My aim in this section is to motivate and to articulate the incoherence objection, to highlight some of its important features, and to distinguish it from several other objections with which it may easily be confused. As I indicated above, I postpone until the next section an affirmative argument for the incoherence objection.

To see the point of the incoherence objection, begin by considering some agent's optimal rules: the rules such that this agent's accepting and sincerely attempting to follow them would promote his interests more than they would be promoted by his accepting and sincerely attempting to follow any other rules; they are the rules that rule egoism would enjoin him to follow. Now clearly it is possible, in principle, for a case to arise in which some agent's optimal rules enjoin him to perform a non-optimal act—an act that promotes his interests less than they would be promoted by some other act that the agent could perform instead.[1] (Indeed we should expect this to be true, since if it were not, then the prescriptions of rule egoism would be entirely consistent with those of egoism[2]—and rule egoism is obviously intended to be an alternative to egoism.) The possibility of such a case is the source of the incoherence objection; this objection regards the possibility of such a case as evidence of a profound *incoherence* in rule egoism. More explicitly, the incoherence objection may be formulated as follows:

> An agent who regards a decision procedure as binding for him solely because he believes that it is optimal for the achievement of some aim deliberates and acts irrationally in a particular case if, in that case, he refrains from performing an act that he believes is optimal with respect to that aim, and opts instead for an act that he believes is non-optimal with respect to that aim, simply because the decision procedure that he regards as optimal for the achievement of that aim prohibits the

[1] Remarkably, this has been denied. Sanders writes, "Rule-egoism does not require self-sacrifice. . . . The 'sacrifices' one makes by refraining from doing the personally optimific thing in the particular situation and doing instead the act enjoined by the operative rule can be recouped in terms of the advantages that accrue through following the rule" (1979, pp. 449–450). He acknowledges that other proponents of rule egoism, such as Hospers, assert the existence of situations in which rule egoism requires self-sacrifice, but he writes, "My point is to urge Hospers to view such situations as involving losses that can be recouped" (p. 450). He does not say how such situations could reasonably be so viewed.

[2] Recall, from chapter I, that in this dissertation 'egoism' refers to the act-based analogue of rule egoism (sometimes called 'act egoism'), not to a family of views of which rule egoism is one variant.

> act that he believes is optimal with respect to that aim. Any theory that prescribes such irrationality in deliberation and action is incoherent.

In stating the objection, it is necessary to index the notion of optimality to "some aim" in order for the objection to be flexible enough apply to a normative theory of any kind. For example, the objection can be understood to refer to rule egoism if we assume that the agent regards a set of rules as binding for him solely because he believes that it is optimal for the maximal advancement of his interests; similarly, the objection can be understood to refer to rule consequentialism if we assume that the agent regards a set of rules as binding for him solely because he believes that it is optimal for the maximal advancement of his society's interests. In the discussion below, references to optimality should be understood as implicitly containing a specification of this sort; when I make this specification explicit, as I sometimes will, I will do so only parenthetically.

The incoherence objection is, of course, familiar. It is known not only by the term 'incoherence', but also by terms such as 'rule-worship' and 'rule-fetishism.'[3] Moreover, considerable space and energy have been devoted to the discussion of this objection; as a result, I am wary of asking the reader to consider it yet again. But certain discussions of it suggest that it is frequently misunderstood or taken insufficiently seriously, and I hope in this chapter to contribute to a solution to these problems.

There are six features of the incoherence objection that I want to highlight. First, it is crucial to understand that in calling certain theories incoherent, the objection is not introducing a stipulative definition of the term 'incoherent': it is not just identifying a certain set of theories and announcing that these are what shall be known herein as incoherent theories. Rather, in calling certain theories incoherent, the objection is drawing on the established meaning—including the derogatory connotation—of the word 'incoherent', and it is saying that we are warranted in attaching this particular label to the theories that prescribe the specified kind of

[3] Other statements of it can be found in Smart (1956, pp. 348–349 and p. 353), Sprigge (1965, pp. 286–287; and 1988, p. 26), Williams (1972, p. 94), Nozick (p. 30), Brandt (1979, p. 296; and 1996, p. 151), Lyons (1980, pp. 25–26), Scheffler (p. 82), Sanders (1988, p. 208), Kagan, (1989, p. 33 and p. 37), Quinton (p. 109), Hooker (1990, p. 74, n. 24; 1994a, p. 95; 1994b, p. 316, n. 8; 1995, p. 28; 1996, pp. 538–539; and 2000, pp. 99–100), Darwall (p. 23), and Blackburn (1998, p. 38). It also seems to be alluded to in Scanlon's claim that "rule utilitarianism . . . strikes most people as an unstable compromise" (p. 103) and in his reference to "the instability which plagues rule utilitarianism" (p. 120). Indeed Hobbes himself seems to anticipate its being deployed against his third law of nature (which is that it is self-interestedly rational to keep certain agreements, even if doing so if involves incurring a net loss), if Kavka is right in claiming that "the rule-worship objection is essentially a restatement of the argument of Hobbes's fool" (1986, p. 378).

irrationality in deliberation and action. Indeed only if it is understood in this way—as a substantive claim of incoherence, not as a stipulation as to how the word 'incoherent' is to be used—can the incoherence objection be understood as identifying a reason for rejecting the theories to which it applies.

A second feature of the incoherence objection that I want to highlight is that the irrationality to which the incoherence objection refers is different from the irrationality to which a theory of rationality refers. An agent exhibits irrationality of the latter sort in deliberation and action if he deliberates and acts in violation of the best theory of rationality (whatever that may be). And insofar as an agent subscribes to and obeys a normative theory (whether a theory of rationality, a theory of morality, or a theory of some other normative domain) that yields prescriptions that diverge from the prescriptions of the best theory of rationality, then he will exhibit irrationality of this latter sort in deliberation and action. But this is not the irrationality to which the incoherence objection refers. In order for the agent to exhibit irrationality of this sort, it is necessary for the agent's deliberations and actions to be plagued by a tension of the specific sort identified by the incoherence objection. And of course this last condition is not only necessary but also sufficient: an agent would exhibit irrationality of this sort even if it were somehow the case that the best theory of rationality were one, such as rule egoism, that necessarily involved the agent in irrationality of this sort.

A third feature of the incoherence objection worth highlighting is that, in order for a theory to be called incoherent by the objection, it is not necessary for the theory to be flatly inconsistent in the sense of prescribing, as rational, deliberation that it (elsewhere, or in some other way) claims or admits is irrational. All that is necessary is for it to prescribe the specified kind of irrationality in deliberation and action, regardless of whether it ever *recognizes* such deliberation and action as irrational. But at this point it should be emphasized that the objection does not accuse of incoherence just *any* theory that prescribes deliberation and action that can reasonably be called irrational. Rather, what makes a theory incoherent, according to the objection, is that it prescribes deliberation and action that are irrational in the particular way specified by the objection.

A fourth crucial feature of the incoherence objection is that the objection may be understood to refer only to those cases in which an indirectly maximizing agent acts *differently* from the way in which a straightforwardly maximizing agent would act. So, if a rule consequentialist is led by her decision procedure to opt for an act that she believes has the best outcome, then the incoherence objection does not arise. To be sure, a critic sympathetic to the incoherence objection might wish to argue that a rule-consequentialist agent *always* deliberates irrationally, just in virtue of deliberating in a rule-consequentialist way, but the incoherence objection does not consist in so sweeping a claim. The incoherence objection is triggered when and only when an indirectly maximizing agent is led by her decision procedure to act in a non-maximizing way: then, and only then, is the agent's deliberation and action irrational in the way that exhibits the incoherence of her decision procedure.

Fifth, the incoherence objection refers only to those cases in which an agent *believes* that he is choosing a non-optimal act. It does not refer to those cases in which the reason for (perhaps I should say 'cause of') an agent's failure to choose the optimal act is that he doesn't notice that the optimal act is open to him, or that he doesn't recognize it as optimal, or some other epistemic incapacity. (It may be claimed that such epistemic incapacity, when it infects agents' deliberations, is a source of irrationality, but I shall not be claiming that in this chapter.) Having said this, I should add that my custom in the remainder of this chapter will be to drop this qualification by making the simplifying assumption that in the cases under discussion, agents are fully informed about their options and their consequences.

A sixth and final feature of the incoherence objection that I want to highlight is that in order to exhibit the sort of irrationality identified by the objection, *the agent's reason* for regarding some set of rules as binding on him *must be* just that it is optimal for some aim. If an agent has some other reason for accepting his optimal set of rules as binding on him, such as (e.g.) that it best summarizes and systematizes his considered moral judgments, then the incoherence objection does not arise (though some other objection might, of course).[4] Another way of putting this point is to say that the agent must follow a certain set of rules that are optimal for some aim *under this description of that set of rules*. So if the agent follows his society's optimal set of rules under the description of it as the set of rules that best summarizes and systematizes his considered moral judgments—with it being merely coincidental that it is his

[4] Hooker notes that he avoids the incoherence objection by justifying the rules of his rule consequentialism in this way rather than in terms of the benefits that accrue to agents who accept or comply with them. He writes, "I'm interested in an argument for RC ['Rule-Consequentialism'] that is *not* founded on an overarching commitment to maximise the good. As I see it, the best argument for RC is that it does a better job than its rivals of matching and tying together our moral intuitions" (1995, pp. 28–29). He later adds, "in denying that the agent need have an overarching commitment to maximise the good, the most plausible versions of rule-consequentialism are repudiating the source of the alleged incoherence" (1996, p. 539). See also Hooker (2000, pp. 101–102 and p. 188).

society's optimal set of rules—then the incoherence objection does not arise. Although this clarification is often omitted from statements of the incoherence objection,[5] it is essential, since acting in compliance with rules that happen to maximally advance certain aims may be justified on grounds quite apart from their relation to those aims.

2.2 There are four objections with which the incoherence objection may be easily confused. The first may come to mind purely on account of the terminological fact that the term 'coherence' has multiple uses in philosophy. Now I assume that not all of the claims that may be expressed using this term are ones with which the incoherence objection may be confused, since many of them concern issues that are very obviously distinct from the one with which we are now concerned. For example, it is probably unlikely for the incoherence objection to be confused with any claim for or against coherence theories of truth or, related to these, coherence theories of the justification of moral theories,[6] since it is probably clear that in accusing indirectly maximizing theories of incoherence I do not mean to be raising any such questions of epistemology, moral or otherwise. But there is another question in connection with which the term 'coherence' is sometimes used, and which may come to mind here because it, like the question addressed in this chapter, arises in debates about the relative merits of act- and rule-based consequentialist normative theories. This is the question, discussed in chapter III, of pragmatic effectiveness, or self-defeat: the extent to which agents' accepting and sincerely attempting to follow the dictates of a particular theory promote or frustrate the achievement of the theory's aims. Lyons, for example, calls the argument claiming that act utilitarianism is pragmatically ineffective "the *general coherence argument*" (1965, p. 151), and Brandt writes that a set of judgments is "causally coherent" when it is "non-self-defeating" (1979, p. 20). Although I might, for reasons set out at the end of chapter III, accuse indirectly maximizing theories of incoherence in the sense suggested by these quotations from Lyons and Brandt, such a claim falls well outside the scope of this chapter. So it is important to distinguish my claim that indirectly maximizing theories are incoherent from the claim that they are pragmatically ineffective or self-defeating in any other way.

[5] Kavka, for example, anticipates this objection and states it simply as, "It is irrational . . . to follow moral rules that generally promote one's interest, in cases in which one believes one can do better by breaking those rules" (1986, p. 378).

[6] See, for example, DePaul (1986 and 1987).

2.3 I hope, then, to have distinguished the incoherence objection from one with which it may be confused for terminological reasons. But there are several other objections with which the incoherence objection may be confused for reasons not of terminology, but of substance—as the literature on the incoherence objection (or purportedly on the incoherence objection) attests. One is the claim that *cases frequently arise* in which an agent stands to gain more (with respect to some aim) from violating the rules that are optimal (with respect to that aim) than from complying with them, or (a similar but not equivalent claim) that cases frequently arise in which an agent is *justified in believing* that he stands to gain more from violating those rules than from complying with them. But the incoherence objection offers no such empirical claim. Rather, it states that *if* such a case were ever to arise—and it does not claim that such a case must or ever will arise—and the agent were to recognize it as such, then the agent would be deliberating and acting irrationally if he were to comply with the rules and (in so doing) to forgo the gains that would be provided by breaking them. So whereas the claim from which I am now distinguishing the incoherence objection is a claim about the actual frequency of certain cases, the incoherence objection presupposes nothing more than the mere possibility of such cases—even if, as things turn out, such cases never arise.

One source of the confusion in which the incoherence objection gets mistaken for this other one may be that Hobbes himself seems to anticipate the incoherence objection in his defense of his third law of nature, as noted above,[7] but then replies, in effect, to this other objection. The objector, whom Hobbes refers to as "the fool" (xv, 4),[8] claims that it is not irrational for an agent to break agreements when doing so is to his advantage. Hobbes's reply, in essence, is that (1) conduct that advances one's interests is not rational unless one could rationally anticipate that it would advance one's interests and (2) breaking one's agreements is likely to alienate one from others whose goodwill one may need in the future (xv, 5). Now the first part of this reply may speak to the incoherence objection, since that objection (as formulated above) refers only to what the agent believes, not to what the agent *reasonably* believes. But clearly, if Hobbes's claim cannot be refuted—and I am not sure that it can't, since it seems quite plausible to say that acts chosen on the basis of irrational beliefs may nonetheless be rationally chosen, or (what comes to basically the same thing) that the rationality of a choice can be

[7] See note 3.

[8] References of the form '*m*, *n*' are to chapter *m*, paragraph *n* of Hobbes.

assessed without assessing the rationality of the beliefs on the basis of which that choice is made[9]—then the incoherence objection could just be restated in terms of what the agent reasonably believes instead of simply in terms of what the agent believes. So the burden of answering the incoherence objection must be borne by the second part of Hobbes's reply: that breaking one's agreements is likely to alienate one from others whose goodwill one may need in the future. But the objection *this* claim answers is clearly not the incoherence objection, but an objection of the sort from which I am now distinguishing the incoherence objection—an objection to the effect that it so often pays to break one's agreements that there's no reason to regard keeping agreements as required by the rules that would be optimal for the advancement of an agent's interests.[10] So whereas Hobbes puts into the mouth of the fool a claim that resonates strongly with readers sympathetic to the incoherence objection, he finds in those words nothing more formidable than the naïve conceit that crime often pays.[11]

[9] Korsgaard, for example, claims that "Judgments of irrationality, whether of belief or action, are, strictly speaking, relative to the subject's beliefs" (1986, p. 318).

[10] Kavka's assessment of Hobbes's reply shows that he, too, reads Hobbes's reply in this way. For he writes: "Hobbes's reply, based on the grave risks of offensive violations . . . succeeds generally, but fails in special cases in which the risks due to violation are both low (compared to the potential gains) and calculable" (1986, p. 378). Kavka could not credit Hobbes's reply with such "genera[l]" success if he thought Hobbes to be replying to the incoherence objection, since the "special cases" in which Kavka concedes that Hobbes's reply "fails" are *all* of the cases to which the incoherence objection refers. But it does make sense to regard Hobbes's reply as a qualified success against the claim from which I am distinguishing the incoherence objection in this subsection.

[11] It appears that Kavka follows Hobbes in mistaking the incoherence objection for this other one, since he says (as quoted in note 3) that the objection of Hobbes's fool is the rule-worship objection (which is the incoherence objection) and (as quoted in note 10) that Hobbes's reply is, for the most part, a success. But once one is mindful of the distinction between the incoherence objection and this other objection, then it is clear that, as we have seen, Hobbes's reply simply fails to answer the incoherence objection by mistaking it for this other one.

Further evidence that Kavka mistakes the incoherence objection for this other one can be found in two other passages in his book on Hobbes (1986). First, on p. 381 (which is in the "Rule Worship" section of his "Rule Egoism" chapter), he claims that "it will not be possible to defend Hobbes's moral theory if it is interpreted as requiring that it be most prudentially rational in every case, for every agent, in every possible (or even actual) social environment to follow the laws of nature and eschew offensive violations" (1986, p. 381). Why Kavka says "requiring" instead of, e.g., "affirming" is unclear. In any case, the evident implication is that if such coincidence *could* be shown, then the incoherence objection would be answered. But the incoherence objection would arise even if such coincidence could be shown, since the following problem would remain: rule egoism would still be committed to saying that if, *per impossible*, a case were to arise in which it were prudentially rational for some agent to commit an offensive violation, then rule egoism would prohibit such an act while referring, for the justification of this prohibition, to the agent's own interests.

Second, on p. 383 (also in the "Rule Worship" section), Kavka refers the reader to a subsequent section of the book for a discussion of one of "Two loose ends of the rule-worship problem." But he concludes this later discussion (1986, pp. 439–446) by writing that "it must be admitted that the attempt to reconcile strong group member's [*sic*] self-interest with a reasonable moral conception of their duties to members of weak groups is less than fully successful. Self-interest, even long-run rational prudence, does not completely explain our obligations to others under all circumstances" (1986, p. 446). But the point of the incoherence objection is not that self-interest cannot completely explain those obligations, but that even if it could, the *possibility* of agents' interests being such that the connection would no longer hold is enough to show the incoherence of a theory such as Hobbes's.

2.4 A third objection from which we must distinguish the incoherence objection is one about what we might call the *motivational adequacy* of indirectly maximizing theories. It might be claimed that once an agent came to see things in the way that motivates the incoherence objection, then he could not possibly be motivated to comply with the rules of the theory in question, and that he could not reasonably be faulted for failing to find in himself the motivation to comply with them.

For an illustration of this possibility, recall (from chapter II) the case of the Humean farmers: my crops will be ready for harvest this week, and yours will be ready next week. Let us say that I am a constrained maximizer, that I sincerely assure you that I will reciprocate if you help me with my crops, and that as a result you do help me with my crops. And let us say that when I am deciding whether to reciprocate, the fact that reciprocating is non-optimal weighs heavily on me—it's awfully hard for me to voluntarily do what I know will make my life go worse, even though I earlier intended to act in this way (as evinced by the fact that I was able to issue a sincere assurance to act in this way) and circumstances are precisely as I anticipated that they would be. Noticing that the only reason for reciprocating that occurs to me is that it is required by my theory of rationality (because, recall, I stand to gain nothing from helping you), I reflect on the justification for that theory. And I recall that this justification appeals to the pragmatic-effectiveness criterion, and to the fact that constrained maximization is the most pragmatically effective theory of rationality that I know of. But then this thought occurs to me: "If what justifies my conception of rationality is that subscribing to it makes my life go as well as possible, then shouldn't I deviate from it when it recommends acting so that my life does not go as well as possible? Shouldn't I deviate from it *now*, for example?" And so I cannot find adequate reason to comply with my theory of rationality; it turns out to be motivationally inadequate.

Now the point of this story may easily be misunderstood, and we need to pause for a moment to focus on the aspect of it that illustrates the motivational-inadequacy objection. One thing that's happening that's *not* very important is that I, as an agent figuring out whether to reciprocate, think that constrained maximization is failing to recommend the act that's truly rational. This fact alone is not very telling since a defender of the theory, such as Gauthier, could just say that my particular judgment about what act is truly rational is simply mistaken, whereas he has an argument—the appeal to the pragmatic-effectiveness criterion—that supports his theory of rationality. The important thing that's happening in the story is that it's precisely *that argument*—that appeal to pragmatic effectiveness—that moves me, as the agent, to abandon constrained maximization in favor of egoism. Even if I didn't have lingering sentiments in support of straightforward maximization, reflecting on the rationale for constrained maximization would drive me to it just the same.[12] Thus could constrained maximization, and any indirectly maximizing theory justified by way of the pragmatic-effectiveness criterion, prove to be motivationally inadequate.

The motivational-adequacy objection is easy to confuse with the incoherence objection because the motivational-adequacy objection follows from the incoherence objection if the latter is supplemented with some additional assumptions about the beliefs and motivations of agents. To see this, notice that the incoherence objection claims that certain forms of deliberation and action are irrational. If we supplement this claim with the assumptions (1) that most agents understand and believe the incoherence objection and (2) that most agents cannot be motivated to deliberate and to act in ways that they regard as irrational,[13] then it can be concluded that a motivation problem arises for any theory for which there can be cases that trigger the incoherence objection. But the incoherence objection does not involve the making of these additional assumptions.

The issue of motivational adequacy seems to be an important one in the eyes of defenders of indirectly maximizing theories. Kavka accords it such importance that the first thing he

[12] Of course, if I (*qua* agent) foresaw that constrained maximization would prove to be motivationally inadequate, then I would be unable to offer the sincere assurance in the first place. As McClennen writes, "If it is expected utility – calculated from the *ex ante* point of deliberation over policies – that judges a CM [constrained maximization] policy to be superior to the SM [straightforward maximization] policy, it is also expected utility – calculated from the *ex post* point of choice of a course of action – that will judge implementation of SM superior to implementation of CM. It would seem, then, that the agent must realize that, notwithstanding the best of intentions, when push comes to shove he will continue to be an SM. . . . The cooperative solution [the one recommended by constrained maximization], then, is simply not feasible" (1988, p. 105; see also p. 106). McClennen concludes, "Gauthier's agent, then, can say to himself and to the other player that he would like to be disposed to cooperate . . . but there is no reason for either to believe that he can be so disposed" (p. 107). See also p. 108: "It appears . . . that he [Gauthier] has only shown that a rational agent would want to be a CM, not that he can be so disposed." A reply is offered by Gauthier (1988).

[13] What are the merits of these assumptions? The first (asserting the widespread acceptance of the incoherence objection) seems, regrettably, to be almost certainly false. The second appears to be more debatable. On the one hand, an agent's belief that some act is rational can, depending on the agent and his circumstances, motivate him to perform it. Brandt, for example, reports that he would be "quite ready to perform the rational act" in a given case if he knew "which it is" (1979, p. 150; see also p. 103). On the other hand, this "mesmeric force"—to borrow a term that Anscombe uses in regard to 'ought', but that seems applicable to 'rational', too (p. 8)—is not always decisive, or even always operative. Unfortunately it would take us far afield to pursue this psychological question to the extent it deserves.

mentions when surveying the merits of rule egoism is that “it has great potential motivational force” (1986, p. 364), and he touts the motivational adequacy of rule egoism in such high-profile passages as the second and second-to-last paragraphs of his book (1986, p. xi and p. 452). Gauthier, too, is concerned enough about the motivational-adequacy problem to criticize Kavka’s solution to it and to propose his own (1987, p. 287). And McClennen, in turn, is concerned enough about the motivational-adequacy problem to criticize Gauthier’s solution to it[14] and to propose one of *his* own (1988, pp. 108–113). But even if one of these solutions succeeds, it is beside the point: the fact that it is psychologically possible for agents to be motivated to deliberate and to act in a certain way, or according to certain principles or deliberative procedures, does not mean that it is *rational* for them to deliberate and to act in that way. Agents can be motivated to deliberate and to act in all sorts of ways, and it may be the case that only one, or a few, of these psychologically possible ways of deliberating and acting is, or are, rational. A solution to the motivation problem is no solution to the incoherence problem.

2.5 The fourth and last objection from which I want to distinguish the incoherence objection is the claim that it is irrational to make oneself into the kind of person who regards, or to dispose oneself to regard, the prescriptions of an indirectly maximizing theory as binding, or as obligatory (whether rationally, morally, or in some other normative way). That is, I want to distinguish the incoherence objection from the claim that it is irrational to make oneself into the kind of person who deliberates and acts in precisely the way that the incoherence objection characterizes as irrational. This claim might be confused with, or at least be thought to accompany, the incoherence objection in the following way. The incoherence objection accuses the indirectly maximizing agent of deliberating and acting irrationally in certain cases. Assuming that it is irrational to dispose oneself ever to deliberate and to act irrationally, the incoherence objection might be taken to imply that it is irrational to dispose oneself to regard the prescriptions of an indirectly maximizing theory as binding.

But this assumption—the assumption that it is irrational to dispose oneself ever to deliberate and to act irrationally—is not a component or even an ally of the incoherence objection. On the contrary, granting this assumption creates problems for the incoherence objection precisely because it causes the incoherence objection to entail the further objection from which I am now distinguishing it. Indeed we will see in section 3 that one strategy for

[14] See note 12.

rebutting the incoherence objection is to insist on a strong link between, on the one hand, the rationality of the adoption and maintenance of certain dispositions and, on the other hand, the rationality of the deliberation and action that flow from those dispositions.

Once we set aside the assumption that it is irrational to dispose oneself ever to deliberate and to act irrationally, then it is clear that the incoherence objection is perfectly compatible with many claims that it may otherwise be thought to contradict. For example, the incoherence objection is compatible with the claim that there are several "sorts of reasons why it would, in general, be rational . . . to adopt . . . a conscientious attitude toward moral rules" (Kavka 1986, p. 382), with the claim that allowing that "adopting or developing a sense of fair play or justice can be entirely consistent with promotion of one's long-term interests" (Kavka 1986, p. 405), and with the claim that "even in terms of self-interest in the narrowest sense, [one's] commitment to morality may be rational" (Kavka 1986, p. 433).[15] The incoherence objection even allows that the acts of self-sacrifice that may result may be "rational in the sense that they are the natural or predictable outcomes of . . . rational prior acts" (Kavka 1986, p. 429). But the incoherence objection denies that rationality in this latter sense is genuine rationality, and it maintains that however rational it may be to adopt and to have (e.g.) a sense of fair play, the fact remains that it is irrational to act on one's sense of fair play when both (1) doing so frustrates one's interests and (2) one's reason for adopting and having a sense of fair play is that doing so promotes one's interests.

Since the incoherence objection allows that it may be rational for an agent to be disposed to regard the prescriptions of an indirectly maximizing theory as binding, it also accommodates certain kinds of evaluations of agents that it seems reasonable to make. For example, the incoherence objection allows that an agent whom it accuses of deliberating and acting

[15] A failure to realize that the incoherence objection is compatible with claims such as these may account for Brandt's allegation that the claim that "rule-utilitarians must be rule-worshippers . . . rest[s] on a confusion" (1979, p. 296). For Brandt's response is to point out that better consequences tend to result when agents regard certain rules as binding than when they try to optimize in every case: "Non-omniscient persons acting to keep promises (of certain kinds) will actually succeed in producing more welfare than those who are aiming at maximizing welfare directly" (1979, p. 297).

The critics of rule-based theories (whether of morality or rationality) against whom reactions such as Brandt's *are* appropriate are those who seem not to see how the optimal rules for advancing an agent's interests could be anything other than the principle of egoism. Sanders, for example, writes that "the best thing for one to do would be to adopt a policy of clever compromise between complying with moral rules and violating them whenever one confidently believed he could get away with such violations (1988, p. 198; see also pp. 200–201). Whether Sanders realizes that his "clever compromise" is essentially a reversion to egoism is unclear. On this point see also Mulholland (p. 545).

irrationally in a particular case may deserve to be praised for evidently having succeeded, in the past, in disposing himself to deliberate and to act in that way (if that prior act, of so disposing himself, advances certain aims more than they are frustrated by his irrational act in the present case). Indeed the incoherence objection allows that we may look with approval upon such an agent, since the successful completion of that earlier project of character modification will probably strike us as more important than the failure of the agent to optimize in this particular case. A display of irrationality in the present case may be evidence of an earlier, and more important, exercise of rationality.

What about an agent who acts rationally in the present case, due to his failure to comply with an indirectly maximizing theory? The incoherence objection allows that such an agent may deserve to be blamed for evidently having failed, in the past, to dispose himself to deliberate and to act in compliance with that theory in the present case. Just as a display of irrationality in the present case may be evidence of an earlier exercise of rationality, so may a display of rationality in the present case evince an earlier failure of rationality. For example, an agent who performs a calculation about outcomes in the present case, instead of simply consulting the relevant rule (assuming that the relevant rule does not itself prescribe calculating about outcomes), evinces an earlier failure to dispose himself not to calculate in such cases. But this does not mean that it would be rational for the agent then to disregard the result of that calculation—as if by doing so he could erase that evidence of his earlier failure. Williams sums up the point nicely:

> If calculation has already been made, and the consequences of breaking the rule are found better than those of keeping it; then certainly no considerations about the disutility of calculation could upset that result. And, indeed, it is very difficult to see how *anything*, for a consistent utilitarian, could upset that result. Whatever the general utility of having a certain rule, if one has actually reached the point of seeing that the utility of breaking it on a certain occasion is greater than that of following it, then surely it would be pure irrationality not to break it? (1972, pp. 93–94)

So the incoherence objection allows that it may be rational dispose oneself to eschew calculation in deference to certain rules (such as the prescriptions of rule egoism or rule consequentialism), even while maintaining that exhibiting that deference may be, itself, irrational.

2.6 In the past four subsections I have distinguished the incoherence objection from four others that may be confused with it: an objection about pragmatic effectiveness, an objection about the frequency with which the cases that trigger the incoherence objection arise, an objection about the motivational adequacy of indirectly maximizing theories, and an objection about the rationality of disposing oneself to regard the prescriptions of indirectly maximizing theories as binding. To conclude this section it is worth noticing how the considerations that give rise to the incoherence objection may—even when an author states them explicitly or appears to have them in plain view—be taken to have an entirely different, and sometimes relatively minor, upshot.

One author observes that "rule-egoism will require us *knowingly to perform actions which are not to our greatest interest*" (Berg, p. 214). The claim, as suggested by the author's italics and confirmed by an examination of the context in which it appears, is the central claim of the passage in which it appears. We have also seen that it is the essence of the incoherence objection. But the incoherence objection is not the objection to rule egoism that the author draws from this claim. Rather, the author simply claims that rule egoism, because it may sometimes require deliberate acts of self-sacrifice, "is not a version of *egoism*," where the author understands 'egoism' to refer to a family of views that might be thought to contain rule egoism as a variant (p. 214).[16] That is, the author's point is a verbal one; his comments on whether it may reasonably be dismissed as "'merely' verbal" (p. 213) need not occupy us. For our purposes it is sufficient to note that even when the observation underlying the incoherence objection is at hand, the objection itself may be overlooked.

3 A defense of the incoherence objection

3.1 In the last section, I motivated and stated the incoherence objection, and I distinguished it from several others for which it may be mistaken. I also noted, in closing, that the incoherence objection is sometimes overlooked even when it is practically in plain view. Nevertheless, it remains true that the incoherence objection is frequently invoked; and yet the frequency with which the incoherence objection is invoked in the literature is matched by the infrequency with which it is explicitly argued for—as if the very familiarity of the incoherence

[16] Remarkably, Sanders replies to Berg's criticism by saying that Berg's worries would be allayed if it could be shown that, as a matter of fact, "the degree to which acts produced by the egoist's optimal dispositions tend to differ from the acts he thinks he ought to do, given his egoism" were small (1979, p. 450)—as if Berg were making an empirical, not a conceptual, point. But in a later paper, Sanders follows Berg, both in seeming to see the incoherence objection and then settling for a terminological point. For he writes, "Given that for the egoist, self-interest is the ultimate value, what reason could an egoist have for following general rules . . . when these conflict with self-interest?" But then, instead of developing the incoherence objection, he just claims that such a position "is ill-suited to any theory deserving the name egoism" (1988, p. 208).

objection makes argument for it superfluous.[17] Indeed it is often presented in the form of a question rather than declaratively asserted,[18] as if the burden of proof is on those who would deny this very familiar objection and an answer is being awaited. But however justified this sense of the burden of proof may be, I would like to sharpen the debate over the incoherence objection by offering an explicit argument in support of it.

My defense of the incoherence objection begins with the thought that there are certain cases in which it is *uncontroversial* that the agents' deliberation and action are irrational. Two examples of such cases are suggested in the quotation from Mill given as the epigraph to this chapter. First, consider a military conflict. (I won't try to be true to the case of the "old-fashioned German tacticians who were vanquished by Napoleon.") Presumably, in warfare as in most other enterprises, there are rules the following of which generally promotes success. But this does not mean that it would be rational for a military tactician to follow those rules in every circumstance to which they could be applied. On the contrary, we would rightly regard a tactician as deliberating and acting irrationally if he could see that the rules' purpose would be frustrated if he were to comply with them, and yet were to comply with them anyway simply because they *generally* serve that purpose. Similarly, we would rightly regard a physician as deliberating and acting irrationally (and possibly inhumanely or murderously) if he followed rules that are generally useful for promoting patients' health, and followed them for that reason, while knowing that in doing so he was damaging his patients' health. Now such cases may be regarded as examples, or as a subset, of the cases to which the incoherence objection refers: namely, as cases in which agents comply with rules that generally serve certain purposes, and do so *because* those rules generally serve those purposes, even though they see that, in the circumstances in which they find themselves, those purposes would be better served by their breaking those rules than by their complying with them.

[17] Essentially the same point is made by McClennen, in his discussion of a principle known as separability, which is closely related to the incoherence objection. McClennen writes, "Quite surprisingly, those who have invoked one or another version of separability have been content to define, rather than defend the principle in question" (1997, pp. 239–240). For descriptions of the separability principle, see McClennen (1988, p. 116; and 1997, p. 229, p. 239, and p. 249), DeHelian and McClennen (p. 328), and Gauthier (1997b, p. 5).

[18] The incoherence objection is stated in the form of a question in at least ten of the twenty-two passages catalogued in the first sentence of note 3: Smart (both); Williams; Nozick, Sanders (1988, p. 208), Hooker (1994b, p. 316, n. 8; 1995, p. 28; 1996, pp. 538–539; and 2000, pp. 99–100); and Darwall (p. 23). Hooker offers a representative example: "If the ultimate goal is the maximization of good, is not it incoherent to follow rules when one knows this will not maximize the good? If rules are merely a means to an end, how can one coherently stick to rules when one knows they will not serve that end in the situation at hand?" (2000, pp. 99–100; cf. 1995, p. 28).

I claim, as I said, that cases such as the military and medical cases are ones in which it is uncontroversial that the agents' deliberation and action are irrational, and that such cases are a subset of the cases to which the incoherence objection refers. I also claim that the cases in this subset are not different in any relevant respect from most (though not all) of the rest of the cases to which the incoherence objection refers. (I grant that in some of these other cases, the irrationality of the agents' deliberation and action may not be *uncontroversial*, but I do not regard this as a relevant difference.) Then, it is trivial to conclude that most of the cases to which the incoherence objection refers are cases in which the agents' deliberation and action are, as that objection claims, irrational. (I defer, until subsections 3.4 and 3.5, extending this conclusion to the *rest* of the cases to which the incoherence objection refers.)

So this part of my defense of the incoherence objection (the part that takes care of most of the cases to which the incoherence objection refers) is a small argument with the following structure. One premise is that certain cases—cases exemplified by the military and medical cases—are cases in which the agents' deliberation and action are irrational. The other premise is that these cases resemble, in all relevant respects, most of the cases to which the incoherence objection refers. To be sure, both of this argument's premises need some support; I shall attempt to provide that in the next two subsections.

3.2 In reply to the first premise, it may be claimed that the agents' deliberation and action in the military and medical cases are not necessarily irrational, since the agents may be seeking, in so acting, to serve purposes other than those that their rules generally serve. For example, the military tactician may want to lose a battle in order to dissuade his superiors from continuing to prosecute a war that he regards as unjust, and the doctor may want to let a patient die because the patient himself wants to die, or for some other humanitarian reason. Now such a reply may point to a need to specify the cases more precisely, so that such additional motives are ruled out. (For example, perhaps the phrase 'for that reason' in the description of the case of the physician, in subsection 3.1, should be replaced with 'solely for that reason'.) But this reply does not touch the incoherence objection as stated in subsection 2.1, since in every case to which the incoherence objection as stated there refers, an agent regards certain rules as binding "solely because" they serve certain purposes and chooses a certain act "simply because" of those rules.

There may be other tempting replies to my claim that the tactician and the physician deliberate and act irrationally. For example, it might be claimed that their deliberations and acts are rational because *most* cases are such that complying with the rules does serve their intended purposes. Or it might be claimed that their deliberations and acts are rational because the agents' faith in the rules' general effectiveness could be perfectly adequate to motivate them to comply with them. Or it might be claimed that their deliberations and acts are rational because it would be rational for them to dispose themselves to regard such rules as binding. But all of these replies, it should be clear, merely confuse the allegation of irrationally that the incoherence objection makes with some other criticism of such agents from which I endeavored, in section 2, to distinguish the incoherence objection. Once a clear understanding is obtained both of the nature of the military and medical cases (as cases that satisfy the conditions that trigger the incoherence objection) and of the content of the incoherence objection, the first premise of the main part of my argument can be seen to be unaffected by any of the replies suggested earlier in this paragraph.

A final reply to my claim that the tactician and the physician deliberate and act irrationally is that deliberation and action are saved from irrationality as long as they are in accordance with the decision procedure that is optimal for the achievement of the agent's aims. Hodgson, for example, writes that the incoherence objection is "completely undermine[d]" by the fact that

> To act upon the act-utilitarian principle would preclude the promotion of certain good consequences, and would probably have worse consequences than would to act upon more specific rules of conduct. Therefore it would not be irrational for a benevolent person to act upon specific rules of conduct rather than upon the act-utilitarian principle. (p. 60)

But this, of course, is essentially an appeal to the pragmatic-effectiveness criterion for evaluating normative theories. We considered the prima facie case for such a criterion in chapter III, and although it will be necessary to conduct an all-things-considered weighing of the benefits of that approach against its costs, we are in the process of exploring its costs in this chapter, and the time for that comprehensive assessment will not come until chapter VI.

3.3 The second, and possibly more controversial, of the premises of the small argument outlined at the end of subsection 3.1 is that the cases discussed above—again, the military and medical cases—are similar in all relevant respects to most of the rest of the cases to which the incoherence objection refers: that is, that there is no feature of these cases that makes the incoherence objection valid in regard to them but not valid in regard to most of the other cases to which it refers.

A couple of replies to this claim may seem promising. First, it may be claimed that there are certain aims of such importance that the decision procedure that is optimal for the advancement of any of those aims must have a different status from—perhaps operate at a higher level of seriousness or commitment than—the military or medical rules. After all, the latter may be easily discounted as rules of thumb, or mere guidelines, and it may seem inappropriate to regard the decision procedure that is optimal for the advancement of the agent's interests, or optimal for the advancement of society's interests, so casually. But the military and medical rules do not need to have an inferior status: depending on the circumstances in which an agent lives his life, such as his occupation and the opportunities and tasks with which he is confronted, the military or medical rules that we are discussing may serve not just the limited purposes of military or medical success, but also the grander purposes of making the agent's life, or the world as a whole, go as well as possible. Obviously the cases under discussion could be construed precisely in this way. And surely the deliberation and action of Mill's military tacticians or physician would not be saved from irrationality if the rules they followed so devotedly were not merely the rules of their professions, but were part of the decision procedure that was optimal for their or their society's interests.

Another reply to my claim (that the military and medical cases are not different in any relevant respect from most of the other cases to which the incoherence objection refers) is that the rules assumed to be operative in the military and medical cases are not formulated precisely enough to be the optimal rules for those circumstances. (If this were true, of course, then these would not be cases to which the incoherence objection refers.) How could it be thought that the rules are not formulated precisely enough? Well, it might be thought that if they were formulated precisely enough, then they would include exceptions to account for whatever elements of the circumstances of the cases in question make them cases in which the rules' purposes are better served by their being broken than by their being obeyed.[19] Now to claim this is essentially to claim that the prescriptions of the optimal rules for a military tactician serve not only generally, but indeed universally and invariably, the purpose of military success; and that the prescriptions of the optimal rules for a physician never conflict with the general injunction 'cure patients'. But

[19] Something like this seems to underlie Brandt's claim, apparently in response to the incoherence objection, that "it is only when the prospect of a loss (or gain) from breach is so serious that it becomes salient in her [the agent's] consciousness that we want to recommend that she reflect on whether the gain/loss in benefit justifies some breach of the rules. A conscience-utilitarianism [which Brandt defends] will recommend that this sort of exception will be built into the basic set of moral motivations" (1996, pp. 151–152).

these are empirical claims that are almost certainly false. Military and medical rules that account for all possible contingencies, so as to admit of no exceptions whatsoever, would almost certainly be too cumbersome to be generally acknowledged and consciously endorsed. Even non-specialists know enough about the vagaries of military and medical practice to realize that there is no comprehensive, discursively available algorithm that results in success in every circumstance. Indeed rules that generally serve a certain purpose—even the rules that optimally serve a certain purpose—almost inevitably conflict, in some cases, with the general aim of serving that purpose. To insist otherwise would be to insist that rule-based normative theories be identical in their prescriptions, or what Lyons calls extensionally equivalent (1965, p. 29), to their act-based analogues—which almost all exponents of rule-based normative theories rightly regard as futile.

3.4 I have argued for two premises: that certain cases, such as the military and medical cases, are ones in which the agents' deliberation and action are irrational, and that such cases resemble, in all relevant respects, most of the cases to which the incoherence objection refers. These imply that the deliberation and action in most of the cases to which the incoherence objection refers are rightly labeled by that objection as irrational. But there are some cases to which the incoherence objection refers that are not covered by the foregoing small argument. I discuss these in this subsection and in the next one.

To bring into view a case that differs from the military and medical cases in a subtle but structurally real way, recall the case of the Humean farmers (introduced in chapter II). In this case, which is essentially a variant of the prisoner's dilemma, the players move sequentially, rather than simultaneously, so that the second player knows the decision of the first player and, if the first player cooperates, has the opportunity to exploit rather than repay his cooperation. If we assume that the second farmer's optimal decision procedure prescribes cooperation (because only by being disposed to cooperate can the second farmer induce the first farmer to cooperate rather than avoid the risk of being exploited) then the second farmer finds himself in a case to which the incoherence objection applies. The second farmer's decision procedure would (in its essentials, at least) be something like the following:

> Instead of calculating with a view to identifying your best option, cooperate if the other agent has cooperated. Cooperate even if (through weakness of will or due to any other cause) you happen to calculate and find that there is more to be gained

> from defecting, and even if someone you trust tells you that there is more to be gained from defecting.

But the corresponding decision procedure for the tactician and the physician, as we have been understanding their cases, needn't be so stringent. It may be something like the following:

> Instead of calculating with a view to identifying your best option, follow the rules in *Warfare for Dummies* or *Medicine for Dummies*. But if you do happen to calculate (through weakness of will or due to any other cause), or if someone you trust tells you that there is more to be gained from breaking the rules, then feel free to do so. Do not reject success just for the sake of fidelity to the rules. They are your servant, not your master.

On this understanding of the cases of the tactician and the physician, these cases differ in an important way from the case of the second Humean farmer. For the tactician and the physician can intend, when adopting their calculation-bypassing rules, to violate them in cases that trigger the incoherence objection (i.e., in cases in which they can see that their aims would be better served by breaking the rules than by complying with them). And even in so intending they can expect to reap the calculation-saving benefits of their rules. In contrast, the second farmer *cannot* intend, when adopting his decision procedure, to violate it in cases that trigger the incoherence objection. For if he were to have such an intention, then he would be unable to sincerely assure the first farmer that he will reciprocate (because he would be intending to exploit the first farmer, if the possibility of doing so occurs to him), and he would have to forgo the benefits that his decision procedure would, if only he would adopt it, make available to him.

To understand more fully the attitude towards their rules that the tactician and the physician must have, note what they can and cannot intend. They *cannot* intend, when adopting their calculation-bypassing rules, to engage in calculations designed to *ascertain* whether a given case is a case that triggers the incoherence objection, since so intending would be incompatible with genuinely adopting their calculation-bypassing rules as binding (though it would be compatible with, say, merely having them at hand as useful rules of thumb that they may turn to at their discretion), and it would be incompatible with their reaping the calculation-saving benefits of their rules. But they can intend, when seeking to reap the benefits of regarding their rules, to deviate from them when they can see that non-optimal action is what their rules prescribe. (These benefits depend, to repeat, only on their intending not to engage in calculations designed to *ascertain* whether a particular case is one in which their rules prescribe acting non-optimally.) In contrast, the second farmer cannot intend to deviate from his rules when he can see that non-optimal action is what his rules prescribe.

But we can easily alter the cases of the tactician and the physician to eliminate this difference. For we can easily suppose that the optimal decision procedure for an agent (such as a tactician or physician) will state that no deviations from a certain set of rules are permitted, not that deviations from those rules are permitted when the agent can see that he stands to gain more by deviating than by obeying. To see this, imagine two decision procedures for the tactician: one permitting him to deviate from a fixed protocol when he can see that it is best to do so, and another not permitting any deviations at all. Even though the first decision procedure allows the agent to deviate only when he can see that it is best to do so, what this means in practice is that the agent will deviate whenever he *believes* that he can see that it is best to do so; and the cases in which he believes that he can see this may include many in which he is mistaken. Nor can the problem necessarily be solved if the deviation-permitting clause is more stringent: allowing the agent to deviate, for example, only when he is *absolutely certain* that he can see that it is best to do so. For even the cases in which he believes *this* may include many when he is mistaken. Indeed the agent may be so constituted that any decision procedure containing a deviation-permitting clause, no matter how stringent, is inferior to a decision procedure absolutely forbidding deviations. Because the optimal decision procedures for the tactician and the physician may require compliance as rigidly as the second farmer's optimal decision procedure does, the military and medical cases may be understood not to differ from, but to resemble, the case of the second farmer in the respect under consideration.

3.5 But there is another way in which the case of the second farmer might be said to differ from the military and medical cases. When the tactician and the physician find themselves in circumstances that trigger the incoherence objection, they find themselves in circumstances whose existence is only incidental to their reasons for being disposed to follow their rules. But when the second farmer finds himself in circumstances that trigger the incoherence objection, he finds himself in circumstances whose existence is essential, not merely incidental, to his reasons for regarding, as binding, the rules that require him to cooperate.

To spell out this thought, let us consider in some detail the farmer's rationale for regarding, as binding, rules that require him to cooperate. He realizes that if he does not regard such rules as binding, then he will be excluded from cooperative schemes such as the one under discussion, but that if he does regard such rules as binding, then he will be able to participate in such schemes. (We assume, of course, that the scheme is advantageous enough for the farmer to want to participate even at the cost of contributing his share—though of course it would be even more advantageous if he could somehow reap the benefit without contributing his share.) So the farmer's rationale for regarding such rules as binding is to gain admission to such cooperative schemes. Notice, though, that when the farmer gains admission to a cooperative scheme from which he would have been excluded but for his regarding rules requiring cooperation as binding, he finds himself in a case that triggers the incoherence objection. What this means is that the farmer regards such rules as binding precisely *in order* to find himself in cases that trigger the incoherence objection. Of course, the farmer does not seek to find himself in such cases *because* they are cases that trigger the incoherence objection; nor does he seek to find himself in such cases in order to take maximal advantage of them, since so intending would deny the farmer access to them in the first place. Rather, he seeks to find himself in such cases because even when he does not take maximal advantage of them, he does better than if he had been denied access to them in the first place. So, some cases that trigger the incoherence objection arise only because the agent regards the rules in question as binding. Such cases are, in effect, nothing less than the intended result of so regarding those rules.

But the military and medical cases are not like this. The military tactician, we have been supposing, regards his rules as binding in order to avoid the costs of lengthy calculations and the risk of error that accompanies such calculations. The same is true of the physician. Neither of them regards his rules as binding in order to produce cases that trigger the incoherence objection; on the contrary, their rules could serve their purposes perfectly well even if neither agent ever found himself in such a case. The tactician and the physician may anticipate the occasional occurrence of cases that trigger the incoherence objection, but for them the existence of such cases is an unfortunate side effect of regarding certain rules as binding. It is not part of the rationale for so regarding them.

This, like the difference discussed in the previous subsection, is a genuine structural difference between the military and medical cases, on one hand, and the agricultural case on the other hand. But is this difference enough to blunt the force of the incoherence objection in the agricultural case? The fact remains that in both types of case, the agents' rationales for regarding their rules as binding is that doing so advances their aims. In each case, the agent reflects on the rules that he may regard as binding and opts for those that, if he regards them as binding, promote his aims the most. The fact that the cases that trigger the incoherence objection figure centrally in the rationale for the Humean farmer's plans but figure only incidentally in the tactician's plans and in the physician's plans does not matter. In each case, the agent has secured the benefits that regarding his rules as binding were meant to secure for him; now he has nothing to lose, and everything to gain, from breaking the rules. If it is irrational for the tactician and the physician to adhere to their rules in cases that trigger the incoherence objection, it must also be irrational for the farmer to do so.

This rejoinder seeks to dismiss the structural difference elaborated above by emphasizing the similarity between the military and medical cases, on the one hand, and the agricultural case, on the other: in each case, the agent adopts and continues to accept his optimal rules in order to achieve his aims. But there may remain some suspicion that the structural difference between the two types of case translates into a difference between the two types of case in regard to the agents' rationality.

To dispel this suspicion, let us repeat the strategy employed in the previous subsection by revising our interpretation of the military and medical cases so that they resemble the agricultural case even in the respect now under consideration. Instead of supposing that the tactician and the physician regard their rules as binding simply or primarily in order to avoid the costs of lengthy and occasionally erroneous calculations, let us suppose that they accept them for strategic reasons resembling those that motivate the second farmer. The key fact about the farmer is that the cases in which his rules require him to act non-optimally arise as the intended result of his regarding those rules as binding, and they would not arise if he did not regard those rules as binding. Now let us imagine the tactician and the physician in similar circumstances. Suppose that, before going into a war with multiple expected battles, the tactician believes the following:

> I do not have much natural talent as a military tactician, and I have even less training in it. But I have been drafted, and I cannot avoid going to war. As I think about how to do as well as I can as a military tactician, it occurs to me that I have to choose one of two approaches. One approach is just to plan to always do what seems best to me. The problem with this is that in any given case, there is a pretty good chance that what seems best to me is going to differ from what a well-trained tactician would choose. So if I plan to always do what seems best to me, my opponents will quickly realize that I do not have much training, and they'll take advantage of that by attacking me especially aggressively. On average, I'll be thoroughly defeated.

> My other option is to regard the rules in *Warfare for Dummies* as binding. If I do that, then my opponents will think that I have more training than I do, and they will fight more cautiously against me. As a result, I will find myself facing outcomes preferable to those I would otherwise face. On average, instead of facing thorough defeat, I'll face a choice between victory and mild defeat. Unfortunately the rules are not fine-tuned enough to always lead me to victory, so my commitment to the rules will lead me to forgo some victories (and of course I could not really regard those rules as binding if I intended to break them in order to seize those victories), so on average I can expect to experience mild defeat rather than victory. But this is better than thorough defeat, so I'll regard those rules as binding instead of always doing what seems best to me.

Now the case of the tactician resembles that of the farmer in the respect currently under consideration. The cases in which his rules require him to act non-optimally arise as the intended result of his regarding those rules as binding, and they would not arise if he did not regard those rules as binding. (If he did not regard those rules as binding, he would not, on average, face a choice between victory and mild defeat, but simply an outcome of thorough defeat.)

We can imagine the physician to be similarly situated. We may suppose, for example, that the physician practices in social circumstances in which most physicians are overconfident quacks, continually inflicting on their patients unheard-of remedies that they're *sure* will work, but that usually don't. In such circumstances a physician may regard a certain set of rules as binding because he knows that doing so instills his quack-fearing patients with confidence and hope, the effects of which far outweigh whatever benefits the physician forgoes when he adheres to the rules on those occasions when they prescribe non-optimal treatment. For a physician so situated, the cases in which his rules require him to act non-optimally arise as the intended result of his regarding those rules as binding, and they would not arise if he did not regard those rules as binding.

With the military and medical cases having been revised to resemble the agricultural case in this last respect,[20] it might be thought that they are no longer cases of irrational deliberation and action. But is this really plausible? Focus on the case of the physician, and suppose that he is deciding to refrain from giving a patient a remedy that he believes is the patient's only chance of surviving. He is deciding in this way because this remedy is not included in his optimal decision procedure—in fact, the use of this remedy is explicitly forbidden by his optimal decision

[20] That is, the military and medical cases have been revised so that in them, as in the agricultural case, the agent regards certain rules as binding precisely in order to find himself in cases that trigger the incoherence objection. Note that in each case the agent may regard certain rules as binding not to cause such cases to arise with certainty, but only to increase the probability such cases' arising.

procedure. (It might seem implausible that the optimal decision procedure would explicitly forbid the use of a certain remedy, as opposed to just not endorsing it. But we may suppose, to fill out the story, that this remedy is so well known among physicians and so often misapplied by them that it is optimal for a decision procedure to explicitly forbid its use rather than just not to endorse it.) Is the physician deliberating and acting rationally in deciding to refrain from providing the remedy, on the grounds that this is the conduct that the optimal decision procedure for him requires?

Of course, such deliberation and action could be rational if the physician were worried that giving this patient this remedy would undermine the confidence that future patients would have in his disposition to adhere to his decision procedure. (Remember, his patients are generally quack-fearing, and they respond much better to physicians whom they see as rigid rule-followers than to ones whom they see as always doing what strikes them as best.) But since (by hypothesis) this is a case that triggers the incoherence objection, it must be understood that any future-oriented benefits, such as the physician's continued fidelity to his optimal decision procedure, are outweighed by the benefits of saving this patient's life. In fact, to ensure that we do not allow ideas of such future-looking benefits to cloud our perception of this case, let us suppose that there are *no* such benefits: this patient is this physician's last patient; the confidence that other physicians' patients' have in *them* will not be affected, because this physician's treatment of this patient will remain a secret, and so on. We suppose, then, that the physician is not worried about the costs of giving this patient this remedy. He regards it as a net benefit, perhaps even as a costless benefit.[21] In such a case it is hard to see how the physician's decision to let his patient die could be rational.

What might the physician say in order to defend the rationality of his decision? First, he might say that letting his patient die is part of an *overall* (e.g., career-long) course of action whose benefits far outweigh costs such as the loss of this patient—and, indeed, whose benefits outweigh its costs more than is the case for any other course of action, making it (in a sense) his

[21] We may, of course, doubt the rationality of the physician's confidence in his remedy. After all, the background of our story is that our physician practices in circumstances in which most physicians are overconfident quacks. So the physician's belief that the aims served by his decision procedure (patients' health) will be best served by violating the prescriptions of that procedure may be irrational. But, as noted in subsection 2.3, the argument of this chapter proceeds on the assumption that either (1) the rationality of deliberation and action does not depend on the rationality of belief or (2) the incoherence objection can be revised to refer to what the agent *reasonably* believes instead of simply to what the agent believes. I also assume that, if necessary, the case under discussion could be recast (albeit more elaborately) so that the relevant beliefs of our physician would be reasonable ones.

best course of action. But in reply we may point out that since letting this patient die is a net loss, any benefits that he mentally associates with this loss must have already accrued, and can be neither redoubled nor forfeited by anything he decides to do now. Second, he might say that his patient would have died much sooner if his commitment to his rules had been casual enough for him to be tempted by the present prospect of saving his patient. But in reply we may agree that he acted rationally when he disposed himself to regard his rules as binding, while insisting that he acts irrationally in continuing to regard them as binding in the present case. Third, he might say that his patient would have died much sooner if his commitment to his rules had been casual enough for him to be moved by claims such as this last one (that although he acted rationally when he disposed himself to regard his rules as binding, he now acts irrationally in continuing to regard them as binding in the present case). But in reply we may agree that he acted rationally in disposing himself to regard such claims as unsound, while insisting that he acts irrationally now in regarding them as unsound.

So there are several replies that the physician might offer in order to reply to our charge of irrationality. But none of them really answers to our original intuition that deciding to let this patient die, with no benefits to be gained thereby, is simply irrational. And given that we modified the physician's case in this way in order to make it resemble, in all relevant respects, the case of the second farmer, we may extend the charge of irrationality to that case as well. The incoherence objection succeeds even against the case that may have seemed most immune to it—giving us license to conclude that the incoherence objection is valid in regard to all of the cases to which it refers.

4 Gauthier's reply: the alleged directedness of constrained maximization[22]

4.1 It will be recalled from subsection 4.2 of chapter III that Gauthier introduces a refinement to constrained maximization; and it may appear that this refinement is specially tailored to respond to the concerns that underlie the incoherence objection. Gauthier writes that in order to be rational, it is not enough that deliberation follow a decision procedure that is *effective* for the realization of some aim; it must also be *directed* to the realization of that aim.

[22] Much of this section is adapted from my "The Toxin and the Tyrant: Two Tests for Gauthier's Theory of Rationality."

For example, deliberation that blindly follows the demands of an omnipotent tyrant may be effective (if, e.g., the tyrant rewards agents who deliberate in that way), but that would not make it rational, since such deliberation would be directed not to the advancement of the agent's aims, but purely to the aim of compliance with the tyrant's demands. This refinement, then, seems to capture much of the intuition that underlies the incoherence objection by imposing on rational deliberation a requirement that we may call the *directedness* requirement.

But what, precisely, is the content of the directedness requirement? In order to satisfy it, must a conception of rational choice disallow (as reasons for acting) *any* fact unrelated to the direct advancement of the agent's aims? Clearly not, for if this were the case, then constrained maximization, which embraces (as reasons for acting) facts unrelated to the direct advancement of the agent's aims—such as facts about the compatibility of acts with plans meeting certain criteria—would *not* satisfy it. And of course Gauthier does not mean for the directedness requirement to rule out constrained maximization. So how is it the case that constrained maximization satisfies the directedness requirement, while some other decision procedures—such as the decision procedure that would enable an agent to placate the tyrant—do not?

Recall that in order to succeed in the realm of the tyrant, I must "take [the tyrant's] directives as reasons for acting in themselves, and independently of how they relate to my concerns." As a result, if I choose conduct only *as if* I regarded—as opposed to *actually* regarding—the tyrant's commands as reasons for acting, then my ends will be frustrated by the tyrant. So the peculiar feature of the decision procedure that would be effective in the realm of the tyrant is that it disallows (as reasons for acting) any facts related to the furtherance of the agent's ends. In contrast to the tyrant-placating deliberator, the constrained maximizer is portrayed by Gauthier as consciously striving—though with some "constraints," of course—to further her ends. For example, the second farmer who has received the first farmer's help reasons that

> if . . . reciprocating leaves me better off than I could have expected to be had I not sincerely assured you of my intention to reciprocate, then it is rational for me to reciprocate, even if some other action would then better realize my objectives. (1988, p. 5)

So constrained maximization is *transparent* to the agent in a way that the deliberation that is effective in the realm of the tyrant is not: while the tyrant-placing deliberator would find his ends frustrated if he were to appreciate and be motivated by the pragmatic effectiveness of his decision procedure, the constrained maximizer may continue to find his decision procedure effective for the furtherance of her ends *even while* she appreciates and is motivated by the pragmatic effectiveness of her decision procedure.

It seems reasonable to say, then, that deliberation that satisfies the directedness requirement is deliberation in which the agent is consciously concerned to advance her aims: she deliberates *about* the advancement of her aims, if not necessarily with the aim of maximally advancing her aims on that particular occasion. So while she is not single-mindedly committed to the advancement of her aims in the way that an egoist or an act consequentialist is, the rational agent eschews blind obedience (of the sort that is effective in the tyrant case) and keeps her ends in view.

4.2 It is plain that Gauthier understands the directedness requirement to be accommodating enough to embrace constrained maximization. But does the constrained maximizer truly keep her ends in view in any real sense? When we consider such an agent, it become clear that she succeeds in cases of assurances, threats, and toxins only to the extent that her deliberation is *not* directed to the furtherance of her ends. She succeeds only because, instead of keeping her ends in view, she gives absolute priority to making sure that she follows through on her assurance, or her threat, or her plan to drink the toxin, even though doing so will be costly. We can count on her to make good on her word in these cases only because we can assume that when it is time for her to act, she will be decisively moved to make good on her word, even though doing so will require her to neglect the pursuit of her ends. For at the time of acting, her ends oppose doing what she chooses to do; she has no end-based reason, no outcome-oriented reason, to do what she chooses to do. If this strikes her as worth doing, it is because she has lost sight, however temporarily, of her ends.

We can appreciate the depth of the constrained maximizer's neglect of her ends by imagining how she might reply if we were to ask her, at the time of decision, why she is choosing to follow through. She might appeal to the future, past, or current benefits of choosing to follow through. First, she might say something about the future benefits of following through, such as the pangs of conscience that would attend defaulting and the reputation effects of defaulting, but any reply along these lines would be a non-starter, since in each case it is stipulated that the benefits of following through are outweighed by the costs: in each case, when all the costs and benefits are counted, following through is unambiguously non-optimal. Second, she might mention the benefits that her having been a constrained maximizer has enabled her to secure in the past: your cooperation in the assurance case, your capitulation in the threat case, a million dollars in the toxin case. But at the time of decision, she has already secured these benefits, and no decision then available to her can put them at risk. Third, she might dwell on the benefits that her having been a constrained maximizer enables her to enjoy in the ongoing present: she might fully admit that following through only frustrates her ends, but she might add that even after she frustrates her ends in this way, she will be doing better than if she had never intended to act this way in the first place. But this reply misses the point. She has already secured the benefits of so intending; they are not at risk. So when she takes these benefits to be reasons for following through, then her deliberation is essentially backward-looking, and not truly directed at the furtherance of her ends. If her deliberation were truly so directed, then what would matter to her would be the fact that following through now is just a deadweight loss, and she would see no reason to follow through.

An objection that naturally arises here is that we are just refusing to see the notion of keeping one's ends in view that Gauthier has in mind in imposing the directedness requirement. For if we insist that deliberation in which the agent keeps her ends in view cannot be backward-looking in the way just discussed, then we are led to think that deliberation so directed is simply that of the egoist or the act consequentialist; and we have already acknowledged that this is not what Gauthier means for his notion of directness to entail. After all, Gauthier proposes the directedness requirement precisely in order to deal with some counterexamples to the pragmatic-effectiveness criterion without thereby being driven to retreat to egoism. To see this clearly, it may be helpful to imagine the positions arrayed along a spectrum. At the poles would be the orthodox position and the unrestricted pragmatic-effectiveness criterion, representing exclusive concern with directedness and with effectiveness, respectively. Gauthier's (refined) position would lie somewhere near the middle, since it originates in a concern with effectiveness but is tempered by some concessions to directedness. With this picture in mind, it is clear that in retreating from the unrestricted pragmatic-effectiveness criterion, Gauthier does not mean to retreat all the way back to the orthodox position, but to stop well short of that.

But this objection presupposes that there is some principled requirement of directedness that entails what Gauthier wants it to entail (the rejection of tyrant-placating deliberation) while not entailing what he does not want it to entail (the rejection of constrained maximization). And the first two paragraphs of this subsection can be read as questioning this presupposition: they can be read as claiming that there is no principled requirement of directedness that restricts the pragmatic-effectiveness criterion without entailing a complete retreat to the orthodox position (which endorses only straightforwardly maximizing theories, such as egoism and act consequentialism). I claim, then, that if Gauthier wants to retreat from the unrestricted pragmatic-effectiveness position, then there is no principled stopping point short of the orthodox position. The compromise position that Gauthier seeks to occupy is just conceptually unavailable.

4.3 What are the implications for the incoherence objection? On the interpretation of the directedness requirement that I have been urging, this requirement—far from missing the point or otherwise failing to respond to the incoherence objection—covers precisely the same ground and commits anyone who would impose it (as Gauthier does in the hope of solidifying constrained maximization's defenses) to the orthodox endorsement of straightforwardly maximizing theories. Of course, it is open to Gauthier to drop the directedness requirement, so that he can continue to reject egoism in favor of constrained maximization. But then he must allow that rationality is, in fact, simply a matter of what pays—not only in the case of the toxin, but also in the case of the tyrant.

5 McClennen's reply: the intrapersonal Pareto optimality of resolute choice

5.1 Because McClennen's theory of resolute choice, like Gauthier's theory of constrained maximization, is an indirectly maximizing theory of rationality, it is vulnerable to the incoherence objection. Here is how one critic, Jean Hampton, makes this point:

> McClennen would have us believe that the reason to cooperate deriving from resolute choice should "win" over the utility-maximizing reason. But recall that he justifies resolute choice by arguing that such choice maximizes utility in the long run. So he grants that expected utility reasoning is the *real* authority, and resolution is a feigned authority—feigned for reasons of utility-maximization.
>
> The second player therefore faces a decision between acting from an authority that is feigned, and acting from the authority that is real, and McClennen is arguing that in this case the former would be more authoritative than the latter (and hence the rightful master of action). (p. 275)

The irrationality of Mill's tactician or physician could be put in similar terms: each faces a decision between acting from an authority that is feigned (i.e., his rules) and acting from an authority that is real (i.e., the reason for those rules).

McClennen, though, has attempted to meet this challenge. In response to Michael Bratman's claim that resolute choice is vulnerable to the incoherence objection (Bratman, p. 13), DeHelian and McClennen write that

> if certain benefits can only be secured by intrapersonal coordination of choice or rule-following, then that consideration will be controlling for a rational decision-maker with respect to any activity undertaken with a view to securing those benefits. . . . [This view] involves no "rule-worship" or "habit-worship" of the sort to which Bratman rightly objects. It prescribes rule-following in the face of an incrementally oriented, obvious recommendation to the contrary, *only* in the class of cases in which an agent who is disposed to resist such incremental reasoning secures gains that are not open to incremental reasoners. (p. 329)

But these last-mentioned cases, which DeHelian and McClennen are concerned to isolate, are precisely the cases to which the incoherence objection refers. So this response, rather than replying in any new way to the incoherence objection as understood here,[23] essentially restates the pragmatic-effectiveness criterion.[24] Indeed it follows, with just one sentence intervening, DeHelian and McClennen's broad claim (cited in note 8 of chapter III) that Bratman's approach "fails the most basic of all pragmatic tests: those who reason in the way he recommends do less well than those who reason differently" (p. 329).

5.2 In a later article (1997), although McClennen continues to appeal to the pragmatic-effectiveness criterion in order to justify resolute choice generally (for references, see the last paragraph of subsection 2.2 of chapter III, and note 9 there), he defends resolute choice against the incoherence objection by developing a more nuanced reply that is stated, but not emphasized, in his earlier book. This reply focuses on the fact that the cases that motivate the development of indirectly maximizing theories are cases of dynamic inconsistency: cases in which an agent would like, at one time, to make a certain choice at a later time, but will want,

[23] That this reply may be successful against *Bratman's* criticisms, in which the incoherence objection is developed less explicitly and possibly with somewhat different content from what I offer in this chapter, is a possibility that I shall not pursue here. My interest in this section is in the manner and extent to which the incoherence objection *as developed in this chapter* can be answered by resources provided or suggested in McClennen's work.

[24] Does McClennen regard it as doing more? He seems to: in a later article, to be considered momentarily, McClennen refers to the articles by Bratman and by DeHelian and him as debating whether his approach "involves 'rule-worshipping'" (1997, p. 242).

at that later time, to make some other choice. One way of describing such cases is to say that in them, the agent's earlier self and the agent's later self have conflicting preferences over the later self's possible acts. Now recall that the main idea of McClennen's argument for resolute choice generally is its pragmatic effectiveness: the fact that an agent who is a resolute chooser can expect his life to go better, overall, than can an agent who is an egoist. McClennen's defense of resolute choice against the incoherence objection is essentially a development of this claim. It is, in effect, the claim that resolute choice can be pragmatically justified to the agent not only in terms of the fact that his life will go better *overall* if he is a resolute chooser than if he is an egoist, but also in terms of the twin claims (1) that life will go better for the agent's *earlier* self if the agent is a resolute choose than if he is an egoist and (2) that life will go better for the agent's *later* self if the agent is a resolute choose than if he is an egoist. So McClennen defends resolute choice against the incoherence objection by *factoring* the general pragmatic-effectiveness claim he makes on behalf of resolute choice into two more-circumscribed pragmatic-effectiveness claims that refer to complementary *parts* of the agent's life.

I shall begin to flesh out this characterization of McClennen's defense of resolute choice against the incoherence objection in the next paragraph, but first I want to offer some further clarification of it. To understand McClennen's defense of resolute choice, imagine two different ways in which a resolute chooser's life can go better, overall, than the life of an identically situated egoist. The possibility to which the previous paragraph alluded is this: life goes better for the resolute chooser's earlier self than for the egoist's earlier self, and better for the resolute chooser's later self than for the egoist's later self, resulting in the resolute chooser's life's going better overall (than the egoist's life). Let us say that when the agent's life goes better in this way, then it goes better in a *Pareto-acceptable* way, since when the agent's life goes better in this way, it does so because life goes better both for the agent's earlier self and for the agent's later self.[25] Another, contrasting, possibility is this: life goes better for the resolute chooser's earlier self than for the egoist's earlier self, and although life goes worse for the resolute chooser's later self than for the egoist's later self, life goes better *enough* for the resolute chooser's earlier self to make it the case that the resolute chooser's life goes better overall (than the egoist's life). Let us

[25] For the relevance of Pareto here, see McClennen (1997, p. 220). The essence of the reply now being developed is found in McClennen's earlier claim that "what forms a natural condition for the self being resolute is its becoming aware that there are benefits to both the present self and the future self that will have to be forgone if the self cannot act resolutely" (1990, p. 212).

say that when the agent's life goes better in this way, then it goes better in a *compensatory* way, since when the agent's life goes better in this way, it does so because the agent's earlier self enjoys benefits that might be thought to compensate for the costs that are imposed on the agent's later self. Now McClennen's defense of resolute choice against the incoherence objection requires that the agent's being a resolute chooser make his life go better in a Pareto-acceptable way, not in a compensatory way. For if the agent's being a resolute chooser made his life go better in a compensatory way, then its overall pragmatic effectiveness could not be *factored* into pragmatic effectiveness that redounds both to the benefit of the agent's earlier self and to the benefit of the agent's later self. Whatever the pragmatic merits of a procedure that makes the agent's life go better in a compensatory way might be, it is crucial, in order to understand McClennen's defense of resolute choice against the incoherence objection, to see that this is *not* the way in which he conceives of resolute choice's resulting in the agent's life's going better overall. Rather, he conceives of resolute choice's resulting in the agent's life's going better overall by making life go better both for the agent's earlier self and for the agent's later self.

Now it may seem unlikely that resolute choice could make life go better for the agent's *later* self, since resolute choice—when its dictates differ from those of egoism—requires the agent's later self to forgo some benefit that egoism would direct the agent's later self to seize. But McClennen's defense of resolute choice against the incoherence objection does not depend on the claim (which McClennen would readily admit is false) that life goes better for the agent's later self if the agent's *later self* is a resolute chooser than if the agent's *later self* is an egoist, *given some choice point at which the agent's later self finds himself*. Rather, McClennen's defense of resolute choice against the incoherence objection depends on the claim that life goes better for the agent's later self if the *agent* is a resolute chooser than if the *agent* is an egoist, *given the difference between the choice points at which the agent's later selves may expect to find themselves*.

Gauthier describes a simple example of a case in which this might happen. Ulysses hopes to sail successfully all the way back to Ithaca, but he is worried about succumbing to the temptation of the Sirens. If he is an egoist, then he may decide to take certain steps: he may decide "to stop up the ears of his mariners, so that they will not want to change course, and to bind himself to the mast, so that he will be unable to respond to the Sirens' song" (1997b, p. 13). But what if Ulysses, having read McClennen's *Rationality and Dynamic Choice*, is a resolute chooser? Then he can form a plan that includes sailing past the Sirens, and he can, when the time comes, choose to sail past the Sirens (Gauthier 1997b, p. 13).[26] Now clearly Ulysses, if he is resolute, will enjoy better prospects overall than he will if he is an egoist, however sophisticated an egoist he might be. So resolute choice is pragmatically effective. But this is not all. It is also pragmatically effective in a Pareto-acceptable way. For if Ulysses is resolute, then not only is his earlier self spared the trouble of binding himself and being bound (for his earlier self must bind himself and stay bound, in anticipation of when he yields the floor to his successor, Ulysses's later self), but also his later self is spared the trouble of being bound and having to get unbound. So both Ulysses's earlier self and his later self do better if Ulysses is a resolute chooser than if he is an egoist (Gauthier 1997b, p. 14).

5.3 Now it might be objected that not all of the cases in which resolute choice directs the agent to make choices that the incoherence objection deems irrational have the structure that McClennen's defense presupposes that they have: that is, it might be objected that not all of them are cases in which—granted that resolute choice makes the agent's life go better overall—it does so in a Pareto-acceptable way rather than in a compensatory way.[27] For example, in the case of the second farmer, in which resolute choice would presumably require reciprocation, the second farmer's later self—the one who does the reciprocating—is not necessarily better off, as a result of the agent's being a resolute chooser, than he would have been if the second farmer had been an egoist. For the benefits of the agent's being a resolute chooser are that he gets the first farmer to help him with his harvest, and these benefits might be entirely consumed by his earlier self. To be sure, they might *not* be entirely consumed by the second farmer's earlier self; if as a result of the first farmer's help the second farmer's earlier self has time to do some chores that would otherwise fall on the later self, or if the later self simply feels better due to the earlier self's not having to endure the exhaustions of harvesting alone, then the second farmer's later self may enjoy some of the benefits of the second farmer's being a resolute chooser. But even if some of the benefits persist in this way until the later self comes into being, they may still not be enough

[26] As Gauthier explains, "In this version of the story, the Sirens do not exert an irresistible compulsion on those who hear their singing; rather, they sing so sweetly that they affect the preferences, or evaluations, formed by their hearers, so that, choosing in a normal way to realize their most favored prospect, they head willingly to their doom" (1997b, p. 13). Resolute choosers are capable of deviating, in choosing, from this "normal way."

[27] Indeed McClennen himself claims only that "*for a certain range of cases*, there is a standpoint [that of intrapersonal Pareto optimality] from which the problem of divergent preferences can be resolved" (1997, p. 220, emphasis added). McClennen may, though, intend to restrict the operation of resolute choice to cases in that range; for more on this, see the next paragraph of my text and the block of McClennen's text I quote there.

to offset the burden, borne entirely by the later self, of helping the first farmer with his harvest. If this were the case, then although the *overall* benefits to the second farmer of his being a resolute chooser would (by hypothesis) exceed the costs of his being a resolute chooser, the benefits to the later self might fall short of the costs for the later self. And then resolute choice would make the second farmer's life go better only in a compensatory way, not in a Pareto-acceptable way.

So it is not clear that McClennen's strategy of factoring the benefits of resolute choice into (net) benefits for the agent's earlier self and (net) benefits for the agent's later self is supported by the facts of every case in which resolute choice needs to be defended against the incoherence objection. But rather than insisting on this criticism, I would like to set it aside for now and to consider the success of McClennen's defense of resolute choice against the incoherence objection in regard to those cases to which McClennen's factoring strategy clearly applies: cases in which resolute choice not only makes the agent's life go better overall, but does so in a Pareto-acceptable way. For McClennen is concerned to emphasize that this is how resolute choice works. He writes,

> To be sure . . . there are versions of being resolute that amount to nothing more than allowing decisions you made earlier to tyrannize over you now [e.g., versions of being resolute that make one's life go better only in a compensatory way, or not at all]. But that is not the model of resolute choice I have been defending. I have argued for a model in which the plan that is taken to be regulative of subsequent choice is one that can be defended from the perspectives both of the time of planning and the time of choice [i.e., is a Pareto-acceptable plan]. That defense turns on the consideration that the kind of coordination over time that planning makes possible economizes on scarce resources that are valued both at the time of planning and at the time of choice. (1997, p. 241)

Let us assume, then, that resolute choice makes the agent's life go better overall, and does so in a Pareto-acceptable way.

How does this feature of resolute choice enable it to be defended against the incoherence objection? First, consider a passage in which McClennen responds directly to this objection. Immediately following the passage just quoted, McClennen writes,

> Consider any plan that satisfies the stated condition [i.e., the condition of intrapersonal Pareto optimality]. That the agent has consequential reasons for subsequently departing from the plan, reasons that can be anticipated at the time of planning itself, cannot be regarded as dispositive, neither at the subsequent time of choice, not at the time of planning. To conclude otherwise is to suppose that an agent is incapable of taking that course of action which is first-best rather than second-best. The argument just developed, then, precludes the kind of argument that Smart developed against rule-utilitarianism, to the effect that it

> involves "rule-worshipping," and thus cannot be reconciled with a thoroughgoing consequentialist perspective. On the account just offered, there is a thoroughgoing consequentialist argument for mastering the art of coordination, and that requires disciplining oneself to the logical constraints of plan-guided choice. (1997, pp. 241–242)

Nor for our purposes—of investigating the significance of resolute choice's making the agent's life go better, overall, in a Pareto-acceptable way—what is remarkable about this passage is that it does not depend on resolute choice's having this feature. McClennen does, of course, focus our attention on those cases in which resolute choice does have this feature, by saying at first "Consider any plan that satisfies the stated condition," but what he says in the rest of the passage does not depend on resolute choice's having this feature. For in the rest of the passage he essentially claims (1) that to regard reasons of certain kinds as dispositive is to resign oneself to a view of rationality on which rational agents have to settle for outcomes that are inferior to those that resolute choosers can enjoy and (2) that this fact answers the rule-worship, or incoherence, objection. And yet these claims apply not only to resolute choice conceived of in a Pareto-acceptable way, but also to resolute choice conceived of in a compensatory way (since even resolute choice conceived of in a compensatory way enables the agent to enjoy outcomes that are superior to those that the straightforwardly optimizing agent must settle for). Indeed the long passage just quoted is, like the passage quoted at the end of subsection 5.1, essentially a restatement of the pragmatic-effectiveness criterion, not a rejoinder to the incoherence objection, or the rule-worship objection, per se.

But is there more to be said? Going beyond the defense McClennen offers of resolute choice, does the Pareto-acceptability of the way in which resolute choice makes an agent's life go better provide the basis for a defense of resolute choice against the incoherence objection that is *better* than the defense that would be available if the way in which it made the agent's life go better were not a Pareto-acceptable one? It might seem to. For one way of expressing the point of the incoherence objection in terms that fit into the foregoing discussion is like this:

> It is irrational for any of an agent's temporally-indexed selves (earlier, later, or otherwise specified) to act non-optimally, if that self's only reason for doing so is to conform to a decision procedure that it regards as optimal—*even if* the self in question is a later self who, in and after so acting, is still doing better, from its own temporally-restricted perspective, than it would be doing if one of its predecessors (one of the agent's earlier selves) had not acted so as to conform to that decision procedure.

Once the incoherence objection is understood in this way, then it may seem less plausible. For the later self can hardly complain that the agent's commitment to the decision procedure in question disadvantages him for the sake of the earlier self; on the contrary, by hypothesis, the agent's commitment to the decision procedure in question benefits *both* the earlier self *and* the later self relative to the prospects they would have faced if the agent had been (say) an egoist. Indeed, if the later self had had the chance to choose which decision procedure (the optimal one or egoism) would be in force prior to and during his tenure, he would have chosen the one in question, even though it would require him to act in a non-optimal way. Since this decision procedure is the one that even the later self would have chosen, it might seem obvious that his compliance with it is rational.

But note that the later self would *not* have chosen it for himself if he could choose egoism for himself while choosing the optimal decision procedure for his predecessor. That is, we have seen that the later self prefers (1) that both he and his predecessor conform to the optimal decision procedure to (2) that both he and his predecessor act egoistically, but note that to both of these he prefers (3) that his predecessor conform to the optimal decision procedure while he acts egoistically.[28] And the choice the later self actually faces is not between the first course of action and the second, but between the first course of action and the third. To be sure, the later self would not have access to so desirable a course of action if the earlier self had anticipated that he would choose it, but that does not affect the fact that he now *does* have access to it. So when we examine the later self's preferences over courses of action, his preferences favor conformity to the decision procedure in question only when we remove one of the options that is actually there.[29] As a result, the attempt to justify the later self's conformity to the decision procedure in

[28] Note also the parallel with the prisoner's dilemma and the free-rider problem: in those cases the representative agent prefers (1) that everyone cooperate to (2) that everyone look out for himself, but to both of these he prefers (3) that everyone except him cooperate while he looks out for his own interests.

[29] It might be thought that conformity to the decision procedure in question could be rationalized from the later self's perspective, even when the third option remains in play, if we take the later self's perspective to be constituted not by his *preferences* over options, but by his judgments of the *rationality* of certain options. For if the later self accepts Gauthier's claim that if it is rational for an agent to make *some* choice on some basis, then it is rational for that agent to make *every* choice on that basis (1997b, p. 21), then the later self must regard the third course of action as the least rational one (since it involves inconsistent bases of choice over time—first the optimal decision procedure, then egoism—in the way that the other two courses of action do not).

But we must not infer too much from this ranking. In regarding the third course of action as the least rational one, the later self is not thereby committed to *also* regarding it as one that it would be irrational for him to *choose now*, given that it is available to him. The later self (indeed, the earlier self, too) may regard it as rational, because advantageous, to choose certain courses of action he regards as irrational (though they may tend to be available only rarely, depending on his theory of rationality). So even if we take the agent's perspective to be terms of the later self's preferences rests on a shift in perspective that itself needs to be justified.[30]

It appears, then, that the Pareto-acceptability of the way in which resolute choice makes an agent's life go better does not provide the basis for a defense of resolute choice against the incoherence objection that is *better* than the defense that would be available if the way in which it made the agent's life go better were not a Pareto-acceptable one. In support of this conclusion, recall (from the last section) the final versions of the military and medical cases—the ones in which the circumstances in which the agents find themselves arise as the intended result of their regarding certain rules as binding, and would not arise if they did not regard those rules as binding. These cases are, in addition to being cases of that kind, *also* cases of intrapersonal Pareto optimality: the later selves do better, even in and after complying with their rules, than they would do if the agent had been an egoist. So those cases do not need to be modified in order to serve as test cases for the intuitive force of intrapersonal Pareto optimality, and all of the considerations showing the agents' irrationality in those cases continue to apply. Of course, it would be possible for one to regard those cases differently in the light of the newly discovered or newly emphasized fact that they exhibit intrapersonal Pareto optimality, but I shall not revisit those cases in such depth here. Rather, having presented arguments against the significance of intrapersonal optimality in the abstract, I shall assume that their application to concrete cases is clear, and I shall spare the reader the rhetorical exercise of making that application explicit.

5.4 This section has been devoted to exploring the possibility that McClennen's theory of resolute choice may have features that enable it to be defended against the incoherence objection better than certain other pragmatically justified theories can. In particular, certain remarks of McClennen's—about the intrapersonal optimality of resolute choice—suggest the possibility that an agent's non-optimal choice can be justified *in terms of the aims the agent then has*, and not just in terms of the aims that the agent has in the course of his life. It turns out, though, that the range of cases in which the dictates of resolute choice differ from those of egoism may have to be radically restricted in order for all of them to be cases in which

constituted by his judgments of rationality rather than by his preferences, his perspective may *not* favor conformity to the decision procedure in question over the third option.

[30] This perspective shift took place, in this section, in the penultimate paragraph of the last subsection. The conclusion just reached is, in effect, an endorsement of the doubt expressed in the first sentence of that paragraph: "it may seem unlikely that resolute choice could make life go better for the agent's *later* self, since resolute choice—when its dictates differ from those of egoism—requires the agent's later self to forgo some benefit that egoism would direct the agent's later self to seize."

intrapersonal Pareto optimality holds. Moreover, even in those cases, the promise of justifying non-optimal choice in terms of the aims the agent then has turns out to be as illusory as it initially sounds. I conclude, then, that the incoherence objection is as strong against resolute choice as against any other indirectly maximizing normative theory.

6 Mill on incoherence: an interpretive question

To begin my defense of the incoherence objection, I referred to a passage from Mill's *A System of Logic* (1843). But elsewhere in that work, and in other works by Mill, some passages can be found in which Mill accords to rules an importance in deliberation that may seem to render Mill's thought a target of, rather than a source of support for, the incoherence objection. Obviously nothing in my defense of the incoherence objection turns on whether I am correct in construing Mill as sympathetic to it, but addressing some of these apparently problematic passages from Mill's work will provide an opportunity for me to elaborate and expand on some claims offered above, in section 2, to the effect that certain plausible thoughts about rational deliberation and action that may seem to conflict with the incoherence objection are actually compatible with it.

Later in the chapter of *A System of Logic* in which Mill's "Germans tacticians" remark appears (in the penultimate paragraph of the entire treatise, in fact), after writing that "the general principle to which all rules of practice ought to conform . . . is that of conduciveness to the happiness of mankind, or rather, of all sentient beings (1843, p. 951 [bk. VI, ch. xii, § 7]), Mill adds the following qualification:

> I do not mean to assert that the promotion of happiness should be itself the end of all actions, or even of all rules of action. It is the justification, and ought to be the controller, of all ends, but it is not itself the sole end. There are many virtuous actions, and even modes of action . . . by which happiness in the particular instance is sacrificed, more pain being produced than pleasure. (1843, p. 952 [bk. VI, ch. xii, § 7])

In this passage, Mill allows that acts of the sort that I have been arguing are irrational—acts that frustrate purposes in terms of which the rules requiring those acts are themselves supposedly justified—may be "virtuous." How can this be reconciled with what Mill says about the rule-governed tacticians and physician, and with the incoherence objection? We can start by noticing what Mill says next:

> But conduct of which this can truly be asserted, admits of justification only because it can be shown that, on the whole, more happiness will exist in the world, if feelings are cultivated which will make people, in certain cases, regardless of happiness. (1843, p. 952 [bk. VI, ch. xii, § 7])

This suggests that the virtuousness of an act is influenced by the virtuousness of the feelings that give rise to it—that an act may inherit virtuousness from feelings in much the same way that an act *cannot* inherit *rationality* from the rules or dispositions that give rise to it.[31] Just as we saw above (in subsection 2.5) that the incoherence objection can be reconciled with regarding certain irrational acts as "rational in the sense that they are the natural or predictable outcomes of . . . rational prior acts" (Kavka 1986, p. 429), so it can be reconciled with regarding certain irrational acts as virtuous.

Another remark from the same paragraph of the *Logic* that may seem to distance Mill from the incoherence objection is the following:

> the cultivation of an ideal nobleness of will and conduct should be to individual human beings an end, to which the specific pursuit either of their own happiness or of that of others (except so far as included in that idea) should, in any case of conflict, give way. (1843, p. 952 [bk. VI, ch. xii, § 7])

The relevance of this remark to the incoherence objection may not be obvious until one recalls Mill's well-known view, expressed in the sentences following the one just quoted, that what makes the "the cultivation of an ideal nobleness of will and conduct" worthwhile is that it contributes to the happiness of the agent and others. So it may seem that Mill is prescribing adherence to that project of cultivation even in cases to which the incoherence objection refers. But it is not clear that Mill is referring to such cases. For one of the recurring themes in Mill's discussions of the importance of rules is that agents are in a position to recognize such a case much more rarely than they tend to realize. As Crisp writes, it is Mill's opinion that

> It is just not clear in practice whether, in any particular case, one might maximize by breaking the rule; and, because it can be assumed that it usually will not maximize to break the rule, breaking the rule should not usually be considered. (p. 118)

[31] Lurking beneath this remark is the thought that although judgments of rationality are primarily of particular events such as acts, choices, and deliberations (with judgments of the rationality of enduring things such as persons and character traits being secondary and derivative), judgments of virtuousness are primarily of enduring things such as persons and character traits (with judgments of the virtuousness of particular events such as acts, choices, and deliberations being secondary and derivative). I do not mean to suggest that this thought is uncontroversial: Gauthier, for example, "presuppose[s] that it is primarily to the individual that we ascribe rationality" (1975a, p. 210). But I do mean to suggest—though I cannot here argue—that such an understanding of these terms is consistent with Mill's use of them.

So when Mill writes that the pursuit of happiness should give way to the cultivation of character, he very well may mean to be making an epistemic point: that the agent ought, for epistemic reasons, to regard himself as not really in a situation in which the pursuit of happiness is in conflict with the cultivation of character. If so, then Mill is not referring to a case that triggers the incoherence objection but is, rather, making an epistemic point not unlike the one we saw, in subsection 2.3, Hobbes making in the first part of his reply to the fool. As we noted there, the incoherence objection can be modified, if necessary, to refer only to cases in which such epistemic worries do not arise. So we need not interpret this remark of Mill's as conflicting with the spirit of the incoherence objection.

One other passage from Mill's work that may seem to accord to rules a more important role in determining the rationality of deliberations and acts than is compatible with the incoherence objection can be found in his *Utilitarianism*. There he replies to the objection that "there is not time, previous to action, for calculating and weighing the effects of any line of conduct on the general happiness" (1861, p. 224 [ch. II, par. 24]) in the following way:

> The proposition that happiness is the end and aim of morality does not mean that no road ought to be laid down to that goal, or that persons going thither should not be advised to take one direction rather than another. . . . Nobody argues that the art of navigation is not founded on astronomy because sailors cannot wait to calculate the Nautical Almanac. Being rational creatures, they go to sea with it already calculated; and all rational creatures go out upon the sea of life with their minds made up on the common questions of right and wrong, as well as on many of the far more difficult questions of wise and foolish. (1861, pp. 224–225 [ch. II, par. 24])

On the basis of this passage it may appear that Mill regards rules as central elements in rational deliberation and action. But as Smart points out, the case of the sailors who use the Nautical Almanac is not representative of the cases with which we are concerned. He writes,

> The example of the nautical almanack is misleading because the information given in the almanack is in all cases the same as the information one would get if one made a long and laborious calculation from the original astronomical data on which the almanack is founded. (p. 350)

This means that, in effect, the optimal rules for sailors—or that subset of their optimal rules that contains rules referring to navigation—never conflict with the goal of correct navigation. So the incoherence objection simply cannot arise. In order to modify the situation of sailors so that it can arise, let us suppose that the almanac is known to have an error rate of one percent (because of occasional disturbances in the apparent positions of the stars, or because of a propensity of the compilers of the almanac to miscalculate occasionally, or for some other reason). But let us also suppose that, as before, sailors can verify the accuracy of entries in the almanac by way of "a long and laborious calculation from the original astronomical data on which the almanac is founded" (Smart, p. 350). Then, as Smart concludes,

> Seafarers might use the almanack because they never had time for the long calculations and they were content with a 99% chance of success in calculating their positions. Would it not be absurd, however, if they did make the correct calculation, and finding that it disagreed with the almanack calculation, nevertheless they ignored it and stuck to the almanack conclusion? (p. 350)

So Mill's claim that rational agents "go out on the sea of life" with their minds made up on many questions does not conflict with the incoherence objection. One can say that Mill's "old-fashioned German tacticians" were rational to go into the fields of battle with their minds made up on certain general rules, but that they still deliberated and acted irrationally when they complied with those rules once they saw that doing so would frustrate the aims for which the rules were selected.[32]

7 Conclusion

As I said earlier in this chapter, the incoherence objection is vastly more often simply mentioned than explicitly argued for by its proponents. And, as I explained earlier in this chapter, it is strangely ignored, or mistaken for some other objection, by surprisingly many advocates of theories that are among the targets of the incoherence objection.[33] (We saw, earlier, that this is unfortunately true of Hobbes.) This absence of discussion of the incoherence objection specifically—as opposed to other objections that have received more attention—prevents me from credibly claiming that I have canvassed and answered a variety of opponents' replies to the incoherence objection. No doubt my argument would be more appealing if I could make such a

[32] Further support for the foregoing interpretation of Mill's thought is provided by Crisp's claim that Mill is an act, not a rule, utilitarian (see, e.g., pp. 102–105).

[33] To be fair, I should acknowledge that Sanders does, if I interpret his work correctly, confront the incoherence objection head-on (1976, p. 274, and 1978, p. 297). But I decline to discuss his reply, since (1) it refers specifically to Hospers's nonstandard version of rule egoism (stated in note 10 of chapter I) and is not clearly adaptable to standard versions of indirectly maximizing theories such as those under consideration in this chapter, (2) it muddies the waters by regarding the incoherence objection as the claim that "rule-egoism *reduces* to act-egoism" (1976, p. 274), and (3) it is sufficiently brief and cryptic to be unsuccessful—and, I think, not illuminatingly so—unless improved by considerable charitable embellishment. There is also the *ad hominem*, but not completely irrelevant, consideration that Sanders ultimately comes to reject rule egoism (1988).

claim. But I can claim to have presented a thorough argument for the incoherence objection and to have addressed several misunderstandings and points of resistance that might interfere with its acceptance.

I have shown, in particular, that a defender of the incoherence objection can allow and even insist that cases almost never arise in which an agent can reasonably expect to advance his aims by breaking rules that are optimal for the advancement of those aims (subsection 2.3), that theories that the incoherence objection targets may be motivationally adequate (subsection 2.4), and that it is not irrational for an agent to dispose himself to comply with the prescriptions of an incoherent theory (subsection 2.5). What a defender of the incoherence objection *cannot* allow is that an agent may deliberate and act rationally in continuing to subscribe to a decision procedure, because of its optimality, in those cases in which directs the agent to settle for a non-optimal outcome (roughly speaking—see subsection 2.1 for the definitive statement). A strong prima facie case for this claim emerges from a careful consideration of a series of hypothetical cases (section 3). Although Gauthier's introduction of his directedness requirement may seem adequate to defend constrained maximization against the incoherence objection, it actually amounts (on the interpretation urged here) to a capitulation to it (section 4). Finally, McClennen's insistence on the intrapersonal Pareto optimality of resolute choice may seem to put that theory on especially strong ground, but this notion turns out not to have any special justificatory force (section 5).

Such, then, is the case for the incoherence objection. But it is, of course, only one consideration of three with which this dissertation is concerned. We have yet to weigh the merits of incoherence objection, and the straightforwardly maximizing theories that it endorses, against those of the pragmatic-effectiveness criterion and the publicity condition, and the indirectly maximizing theories that they seem to endorse. It is this perspective that we shall take up in the next, and final, chapter.

VI

Conclusion

1 Theories revisited

1.1 The last three chapters have been devoted to examining the content and validity of three standards commonly used to evaluate normative theories: that a theory must not be self-defeating (in the sense of being pragmatically ineffective), that a theory must not violate a publicity condition of some form, and that a theory must not be incoherent in a certain way. We found that although the first two standards are vulnerable to serious criticisms (chapters III and IV), the third may be defended against the replies that may be offered against it (chapter V).

The point of examining these standards, it will be recalled, is to gain a better understanding of the merits of the theories that these standards are commonly used to criticize. And since the self-defeat and publicity standards are commonly used to criticize straightforwardly maximizing theories, while indirectly maximizing theories are vulnerable to the incoherence objection, the results just described might appear to provide unequivocal support for straightforwardly maximizing theories against their indirectly maximizing rivals.

But we should not move from judgments about standards to judgments about theories so quickly. For each of the judgments about standards at which we arrived was based on a balancing of intuitions for and against the standard in question, considered in isolation from the others. Each judgment was, in effect, an 'other things being equal' judgment. And yet it could turn out to be the case that when all of the standards are considered together, the self-defeat standard and the publicity condition overpower the concern about incoherence in that indirectly maximizing theories—even if the standards in support of them seem vulnerable to criticism when examined one by one—could turn out to be more intuitively compelling, on the whole, than straightforwardly maximizing theories.[1] It is this possibility to which we now turn.

[1] This point can be illustrated by way of a simple example from physics. Suppose a car is on a mild slope. Then we would expect each of the following to be true, other things being equal (i.e., taking each statement in isolation from the others): (1) I cannot push the car uphill, (2) you cannot push the car uphill, and (3) the brake's being off would be sufficient to make it go downhill. Each judgment, taken by itself, suggests that the car will not

1.2 We cannot, of course, explore this possibility in the depth it deserves. But we can bring together some of the results of the previous chapters. In particular, since straightforwardly maximizing theories tend to serve as the default positions to which indirectly maximizing theories are offered as improvements, we can survey the ways in which the latter theories are thought to be superior to the orthodox theories they are designed to displace, and weigh those alleged improvements against the demerits of departing from their orthodox rivals.

What, then, are the supposed advantages of indirectly maximizing theories over straightforwardly maximizing ones? First, as the results of chapter II suggest, they tend to be more pragmatically effective, meaning that they may be not be self-defeating in the way that straightforwardly maximizing theories are. Now I argued in section 6 of chapter III that certain indirectly maximizing theories—rule-based ones such as rule consequentialism and rule egoism—may be far less pragmatically effective than it is often supposed. But other indirectly maximizing theories—non-rule-based ones such as constrained maximization and resolute choice—may not have this problem.

But what is the value of avoiding self-defeat? We saw in chapter III that although a theory's being pragmatically effective may have some initial intuitive appeal, the ultimate justification of a theory cannot be pragmatic; rather, the pragmatic-effectiveness criterion itself (the self-defeat standard) must be justified in theoretical terms—in terms of the intuitions about rational and moral action that it embraces but that the orthodox approach (with its endorsement of straightforwardly maximizing theories) must reject.[2] And we saw that taking a pragmatic approach to evaluating normative theories saves several attractive intuitions that are lost with straightforwardly maximizing views. A fairly concrete one is that it is rational for an agent to choose acts that it is rational for that agent to dispose herself, or to have disposed herself, to choose. A somewhat more abstract one is that well-reasoned action—whether rational or moral—is pragmatically effective.[3] McClennen, for example, characterizes his defense of resolute choice as "a brief for rationality as a positive capacity, not a liability – as it must be

go uphill. And yet if we consider all of the factors together—summing, in effect, the *forces* instead of the *judgments*—we might think that the car will go uphill, if we think that you and I can push hard enough to counteract the force of gravity.

[2] So we must dissent from McClennen's remark that "within an adequate theory of rational choice, intuition must give way to well-grounded pragmatic arguments" (1997, pp. 240–241). For on our view, intuitions do not *give way* to well-grounded pragmatic arguments; rather, they *ground* them.

[3] Note the role of this thought in a theoretical, not pragmatic, justification of indirectly maximizing theories.

on the standard account" (1988, p. 118). Finally, as we saw that Gauthier emphasizes, an indirectly maximizing agent exhibits a kind and degree of self-mastery that it seems reasonable for humans to aspire to, and that is captured in Nietzsche's rhapsodic praise of people who can "stand as their own guarantors."

These, then, are some of the intuitions that count in favor of indirectly maximizing theories in virtue of their avoidance of self-defeat. Now it might also be claimed on behalf on such theories that they tend not to violate the publicity condition in the way that straightforwardly maximizing theories do; but we saw in chapter IV that even indirectly maximizing theories—indeed, all except some terribly implausible normative theories—run afoul of the publicity condition. So it seems fair to say that the principal advantages of indirectly maximizing theories over straightforwardly maximizing ones are due to their affinity with the pragmatic-effectiveness criterion.

1.3 But do these benefits outweigh the losses that must be endured in order to pay for them? The central problem with indirectly maximizing theories, as we saw in chapter V, is that they suffer from a serious, fundamental, and ineradicable incoherence in virtue of requiring agents to forgo identifiable and attainable benefits for the sake of compliance with a decision procedure whose justification is nothing more than that the acceptance of such a decision procedure tends to lead to better results than the acceptance of some other (e.g., a straightforwardly maximizing) decision procedure. And recall that this problem arises even in those cases in which there is no doubt about the availability of the benefit to be obtained by the agent's departing from her decision procedure, about the reasonableness of the agent's belief that so departing will result in the obtaining of the benefit, or about whether the benefit outweighs the long-term costs of the agent's departing from her decision procedure. Theories with this problem—theories that declare that consequentialist reasons really require agents to forgo outcomes with the best consequences—may reasonably be regarded as seriously flawed.

Moreover, the damage that indirectly maximizing theories suffer from their incoherence may overwhelm the advantages already surveyed. For the intuitions that such theories save in virtue of their pragmatic effectiveness—about the rationality of executing rationally formed plans, about the pragmatic effectiveness of rational and moral thought, and about the desirability of the self-mastery that indirectly maximizing agents may be able to exhibit—are hardly necessary truths. And the claims that straightforwardly maximizing theories require us to substitute for these are hardly untenable. First, once we see that it can be rational to make plans for reasons that go above and beyond the reasons that would support their execution (e.g., for the autonomous effects of the intentions thus formed), we can see how it might not be rational to execute a plan that it was rational to make. Second, McClennen is surely right to say that there is something "paradoxical" about the position, which must be owned by the defender of straightforwardly maximizing theories, that

> a fully rational person, faced with making decisions over time, will do less well in terms of the promotion of various standard values than one who is capable of a special sort of "irrationality." (1997, p. 240)

But such a position needn't be regarded a more problematic than, say, the position that false beliefs may sometimes be more useful for the attainment of true ones than other true ones will be. Third, in response to the Nietzschean ideal of the self-mastering agent, it may be said that standing as one's own guarantor, though it certain sounds appealing, actually involves allowing one's present choices to be determined by past wishes and is thereby incompatible with the sort of full, ongoing, conscious control of one's choices and actions that well-reasoned deliberation and action ultimately require.

These, then, are some of the intuitions that must be taken into account. To be sure, there are others.[4] But the ones discussed above seem likely to be the most important. In any case, intuitions such as these must ultimately decide the issues raised in this dissertation. As John Heil writes (in a completely different context),

> Philosophers, lest we forget, can ill-afford to eschew platitudes. Although such things are by no means infallible, they comprise whatever bedrock there is for our speculations. (1984, p. 59)

Obviously different philosophers will find some areas of bedrock more solid than others. But I hope to have identified the regions, and the salient features, of the bedrock on which any sound theory of morality or rationality must be built.

[4] For example, it may be said on behalf of indirectly maximizing theories of rationality, such as constrained maximization, that they enable rational agents' to *trust* one another in a way that straightforwardly maximizing theories of rationality cannot. That is, they happily dissent from the egoist's claim that it is irrational to trust a rational person in the sense of relying on him to do something that he does not have a net incentive to do. But the significance of this point has to be assessed in the light of how few occasions for such trust actually arise. For it has long been recognized that one of the keys to a well-functioning society is to minimize the frequency of such occasions by keeping records that lead to the accumulation of reputations that agents are loathe to ruin. As Axelrod writes, "The foundation of cooperation is not really trust, but the durability of the relationship. . . . Whether the players trust each other or not is less important in the long run than whether the conditions are ripe for them to build a stable pattern of cooperation with each other" (p. 182).

2 Impossibility

2.1 One natural response to the debate canvassed in the previous section—between straightforwardly maximizing theories and their indirectly maximizing rivals—is to ask whether one might construct a theory that combines the merits of the ones considered up to this point. Can there be a theory that avoids the problem of self-defeat (and, we might add, satisfies some plausible version of the publicity condition, if there is one), while avoiding incoherence?

Unfortunately the prospects for such a theory—at least within the confines of plausible forms of consequentialism about morality and rationality—seem dim. One way of constructing a non-self-defeating, non-incoherent consequentialist theory is to specify the good in such a way that it is so easy to achieve that problems of self-defeat do not arise. (Recall, from chapter III, the theory of rationality on which an act is rational if and only if it minimizes the agent's exposure to the sun. Such a theory obviously has no incoherence problem; and it would appear to be so easy to implement that problems of self-defeat would not arise.) But any specification of the good that is simple enough to avoid problems of self-defeat is also likely to be too simple to be plausible. (Avoiding exposure to the sun hardly seems like a plausible specification of the good for a person.) That is, once the good is recognized to include anything as complex as pleasures and pains—not to mention desire-satisfaction, the pursuit of rational plans of life, and the development of virtuous character traits—problems of self-defeat appear to be inevitable for any straightforwardly maximizing theory. It may, then, be impossible to construct a theory that is both non-self-defeating and non-incoherent.[5]

2.2 Now one response to the foregoing impossibility claim is just to say, "So much the worse for consequentialism: if every consequentialist normative theory is (1) self-defeating, (2) incoherent, or (3) implausible for other reasons, then that's a *reductio ad absurdum* of

[5] Impossibility results, though rarely stated as such in moral philosophy, are found from time to time in other fields. Perhaps the most famous impossibility result is Arrow's theorem, which shows that no social welfare function can satisfy all of several conditions that it seems reasonable to impose on such functions.

Another, less technical, impossibility result has to do with geographic maps. We might think, for example, that any acceptable map will satisfy at least the following two basic and seemingly innocuous requirements: that it correctly represent the relative areas of land masses, bodies of water, and so on; and that it enable a traveler (such as a sailor or aviator) to find the direction in which to travel—in order to get from one point to another—simply by drawing a straight line connecting the two points. As it turns out, though, no map can satisfy both of these requirements. The first is met by the Peters projection, and the second is met by the Mercator projection, but neither map meets both of them, and none can (Monmonier, pp. 13–19). This example shows, in addition, that 'impossibility result' is just a fancy term for not being able to have everything one wants in a theory, function, map, or other construct.

consequentialist normative theorizing itself." An inference of this kind is sometimes made in response to Arrow's theorem. As Elster writes of the difficulties Arrow's theorem highlights, "To some extent the difficulties also point to objections, since the various impossibility theorems of the theory also provide reasons for thinking that something must be wrong with the whole framework" (1983, p. 31).

But this response to the impossibility result suggested above is a rather extreme one. For it seems reasonable to admit that no consequentialist normative theory is capable of having all of the features that we would like for a normative theory to have, while maintaining that any nonconsequentialist theory will be (for reasons not discussed in this dissertation, obviously) *ipso facto* untenable. Such a position would not require a dogmatic commitment to consequentialism; on the contrary, it could be the result of a careful weighing of the pros and cons of consequentialist and non-consequentialist theories, not unlike the weighing of the pros and cons of straightforwardly and indirectly maximizing (consequentiaist) theories undertaken in the last section. On such a basis one might conclude about some consequentialist moral theory, such as utilitarianism, "that it is a very bad form of moral philosophy, but that all the others are so much worse" (Scarre, pp. 1–2); and one might conclude something similar about some consequentialist theory of rationality.

2.3 On this view, the proper response to the impossibility result suggested above would be to realize that we are simply asking too much of a consequentialist theory if we require of it that it be neither self-defeating nor incoherent. That is simply not something that consequentialism, in any plausible form, can be. Such a response would involve, moreover, recognizing the inadequacy of two common argumentative moves: first, the rejection of certain consequentialist theories as self-defeating; and, second, the rejection of certain consequentialist theories as incoherent. For if the impossibility result suggested above could be established, it would show the inadequacy—we might say the *incompleteness*—of each of these two moves (insofar as it was offered as sufficient to dispose of the theory in question). For if every plausible consequentialist theory is either self-defeating or incoherent, then to show *which* of these any given consequentialist theory *is* tells only half the story. The rest of the story, which normative theorists typically neglect but which very much needs to be told, is this: Assuming (as seems evident) that each of these two standards has some intuitive appeal, on what reasonable basis can one take precedence over the other? Which is truly the more compelling? Reorienting the debate in this way would lay bare the issues underlying the terms in which the debate is currently conducted, and might allow consequentialist theorizing to escape the impasse, in which it now finds itself, created by the current debates over self-defeat, the publicity condition, and incoherence.

BIBLIOGRAPHY

Bibliography

Adams, R. M., “Motive Utilitarianism,” *The Journal of Philosophy* vol. 73, no. 4 (August 12, 1976), pp. 467–481.

Anscombe, G. E. M., “Modern Moral Philosophy,” *Philosophy* vol. 33, no. 124 (January 1958), pp. 1–19.

Aristotle, *Nichomachean Ethics*, trans. by W. D. Ross, rev. by J. O. Urmson, in *The Complete Works of Aristotle*, ed. by Jonathan Barnes, vol. 2 (Princeton: Princeton University Press, 1984), pp. 1729–1867.

Arrow, Kenneth J., *Social Choice and Individual Welfare*, second ed. (New Haven, Conn.: Yale University Press, 1963).

Audi, Robert (ed.), *The Cambridge Dictionary of Philosophy*, second ed. (Cambridge: Cambridge University Press, 1999).

Axelrod, Robert, *The Evolution of Cooperation* (New York: Basic Books, 1984).

Baier, Kurt (1965), *The Moral Point of View: A Rational Basis of Ethics*, abridged edition (New York: Random House, 1965).

——— (1995), *The Rational and the Moral Order: The Social Roots of Reason and Morality* (Chicago: Open Court Publishing Company, 1995).

——— (1996), “Comments,” in Schneewind, pp. 210–286.

Barnes, Gerald, “Utilitarianisms,” *Ethics* vol. 82, no. 1 (October 1971), pp. 56–64.

Bentham, Jeremy, *An Introduction to the Principles of Morals and Legislation*, ed. by J. H. Burns and H. L. A. Hart (Oxford: Clarendon Press, 1996).

Berg, Robert, “Rule-Egoism?” *The Personalist* vol. 60, no. 2 (April 1979), pp. 211–215.

Bertram, Christopher, “Self-Effacing Hobbesianism,” *Proceedings of the Aristotelian Society* vol. 94 (The Aristotelian Society, 1994), pp. 19–33.

Blackburn, Simon (1994), *The Oxford Dictionary of Philosophy* (Oxford: Oxford University Press, 1994).

——— (1998), *Ruling Passions: A Theory of Practical Reasoning* (Oxford: Clarendon Press, 1998).

Blinder, Alan S., *Central Banking in Theory and Practice* (Cambridge: MIT Press, 1998).

Brandt, Richard B. (1967), "Some Merits of One Form of Rule Utilitarianism," *University of Colorado Studies, Series in Philosophy no. 3: The Concept of Morality* (Boulder, Colorado: University of Colorado Press, January 1967), pp. 39–65.

——— (1972), "Rationality, Egoism, and Morality," *The Journal of Philosophy* vol. 69, no. 20 (November 9, 1972), pp. 681–697.

——— (1979), *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979).

——— (1996), *Facts, Values, and Morality* (Cambridge: Cambridge University Press, 1996).

Bratman, Michael, "Planning and the Stability of Intention," *Minds and Machines* vol. 2 (1992), pp. 1–16.

Brink, David O., "Utilitarian Morality and the Personal Point of View," *The Journal of Philosophy* vol. 83, no. 8 (August 1986), pp. 417–438.

Broome, John, review of McClennen 1990, in *Ethics* vol. 102, no. 3 (April 1992), pp. 666–668.

Carson, Rachel, *Silent Spring* (Boston: Houghton Mifflin Company, 1963).

Castañeda, Hector-Neri, "Ought, Value, and Utilitarianism," *American Philosophical Quarterly* vol. 6, no. 4 (October 1969), pp. 257–275.

Cohen, Joshua, "Reflections on Rousseau: Autonomy and Democracy," *Philosophy & Public Affairs* vol. 15, no. 3 (Summer 1986), pp. 275–297.

Crisp, Roger, *Routledge Philosophy GuideBook to Mill on Utilitarianism* (London: Routledge, 1997).

Garrett Cullity and Berys Gaut (eds.), *Ethics and Practical Reason* (Oxford: Clarendon Press, 1997).

Darwall, Stephen, "Reason, Norm, and Value," in Schneewind, pp. 20–38.

Dean, Richard, "A Defence of Constrained Maximization," *Dialogue* vol. 36, no. 3 (Summer 1997), pp. 453–467.

DeHelian, Laura, and Edward F. McClennen, "Planning and the Stability of Intention: A Comment," *Minds and Machines* vol. 3 (1993), pp. 319–333.

DePaul, Michael R. (1986), "Reflective Equilibrium and Foundationalism," *American Philosophical Quarterly* vol. 23, no. 1 (January 1986), pp. 59–69.

——— (1987), “Two Conceptions of Coherence Methods in Ethics,” *Mind* new series vol. 96, no. 384 (October 1987), pp. 463–481.

Eggleston, Ben, “The Toxin and the Tyrant: Two Tests for Gauthier’s Theory of Rationality,” in Kenneth F. T. Cust (ed.), *Twentieth-Century Values* (University Press of America, forthcoming).

Eliot, George, *Romola*, ed. by Dorothea Barrett (London: Penguin Books, 1996; first published in 1863.

Elster, Jon (1983), *Sour Grapes: Studies in the Subversion of Rationality* (Cambridge: Cambridge University Press, 1983).

——— (1984), *Ulysses and the Sirens: Studies in Rationality and Irrationality*, revised ed. (Cambridge: Cambridge University Press, 1984).

——— (1989), *Nuts and Bolts for the Social Sciences* (Cambridge: Cambridge University Press, 1989).

——— (2000), *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints* (Cambridge: Cambridge University Press, 2000).

Gaut, Berys, “Consequentialism,” in Audi, pp. 176–177.

Gauthier, David (1975a), “Reason and Maximization,” in Gauthier 1990b, pp. 209–233; originally published in *Canadian Journal of Philosophy*, vol. 4, no. 3 (March 1975), pp. 411–433.

——— (1975b), “Coordination,” in Gauthier 1990b, pp. 274–297; originally published in *Dialogue* vol. 14, no. 2 (June 1975), pp. 195–221.

——— (1984a), “The Incompleat Egoist,” in Gauthier 1990b, pp. 234–273; originally published in Sterling M. McMurrin (ed.), *The Tanner Lectures on Human Values*, vol. 5 (Salt Lake City: University of Utah Press, 1984), pp. 67–119.

——— (1984b), “Deterrence, Maximization, and Rationality,” in Gauthier 1990b, pp. 298–321; originally published in *Ethics* vol. 94, no. 3 (April 1984), pp. 474–495.

——— (1984c), “Responses to the Paradox of Deterrence: Afterthoughts,” in Douglas MacLean (ed.), *The Security Gamble: Deterrence Dilemmas in the Nuclear Age* (Totowa, New Jersey: Rowman and Allanheld, 1984), pp. 159–161.

——— (1986), *Morals by Agreement* (Oxford: Clarendon Press, 1986).

——— (1987), “Taming Leviathan,” *Philosophy & Public Affairs* vol. 16, no. 3 (Summer 1987), pp. 280–298.

——— (1988), “Morality, Rational Choice, and Semantic Representation: A Reply to My Critics,” *Social Philosophy & Policy* vol. 5, no. 2 (Spring 1988), pp. 173–221.

——— (1990a), “Introduction,” in Gauthier 1990b, pp. 1–8.

——— (1990b), *Moral Dealing: Contract, Ethics, and Reason* (Ithaca: Cornell University Press, 1990).

——— (1994), “Assure and Threaten,” *Ethics* vol. 104, no. 4 (July 1994), pp. 690–721.

——— (1996), “Commitment and Choice: An Essay on the Rationality of Plans,” in Francisco Farina, Frank Hahn, and Stephano Vannucci (eds.), *Ethics, Rationality, and Economic Behaviour* (Oxford, Clarendon Press, 1996), pp. 217–243.

——— (1997a), “Rationality and The Rational Aim,” in Jonathan Dancy (ed.), *Reading Parfit* (Oxford: Blackwell Publishers, 1997), pp. 24–41.

——— (1997b), “Resolute Choice and Rational Deliberation: A Critique and a Defense,” *Noûs* vol. 31, no. 1 (March 1997), pp. 1–25.

——— (1998), “Rethinking the Toxin Puzzle,” in Jules L. Coleman and Christopher W. Morris (eds.), *Rational Commitment and Social Justice: Essays for Gregory Kavka* (Cambridge: Cambridge University Press, 1998), pp. 47–58.

——— (2001), conversation on June 13, 2001.

Gibbard, Allan, “Rule-Utilitarianism: Merely an Illusory Alternative?” *Australasian Journal of Philosophy* vol. 43, no. 2 (August 1965), pp. 211–20.

Griffin, James (1986), *Well-Being: Its Meaning, Measurement and Moral Importance* (Oxford: Clarendon Press, 1986).

——— (1992), “The Human Good and the Ambitions of Consequentialism,” *Social Philosophy & Policy* vol. 9, no. 2 (Summer 1992), pp. 118–132.

——— (1995), “Consequentialism,” in Honderich, pp. 154–156.

Haldeman, H. R., with Joseph DiMona, *The Ends of Power* (Times Books, 1978).

Hampton, Jean, review of McClennen 1990, in *Canadian Philosophical Reviews* vol. 11, no. 4 (August 1991), pp. 273–275.

Hare, R. M. (1963), *Freedom and Reason* (Oxford: Clarendon Press, 1963).

——— (1971a), “The Argument from Received Opinion,” in Hare 1971b, pp. 117–135.

——— (1971b), *Essays on Philosophical Method* (Berkeley: University of California Press, 1971).

——— (1973), “Principles,” in Hare 1989, pp. 49–65 (originally published in *Proceedings of the Aristotelian Society* new series vol. 73 [1972–73], pp. 1–18).

——— (1976), “Ethical Theory and Utilitarianism,” in Hare 1989, pp. 212–230 (originally published in H. D. Lewis, pp. 113–131).

——— (1979), “Utilitarianism and the Vicarious Affects,” in Hare 1989, pp. 231–244 (originally published in Sosa, pp. 141–155).

——— (1981), *Moral Thinking: Its Levels, Method, and Point* (Oxford: Clarendon Press, 1981).

——— (1989), *Essays in Ethical Theory* (Oxford: Clarendon Press, 1989).

——— (1997), *Sorting Out Ethics* (Oxford: Clarendon Press, 1997).

Harman, Gilbert, *The Nature of Morality: An Introduction to Ethics* (New York: Oxford University Press, 1977).

Heil, John (1983a), “Doxastic Agency,” *Philosophical Studies* vol. 43, no. 3 (May 1983), pp. 355–364.

——— (1983b), “Believing What One Ought,” *The Journal of Philosophy* vol. 80, no. 11 (November 1983), pp. 752–765.

——— (1984), “Doxastic Incontinence,” *Mind* new series vol. 93, no. 369 (January 1984), pp. 56–70.

——— (1992), “Believing Reasonably,” *Noûs* vol. 26, no. 1 (March 1992), pp. 47–61.

Hobbes, Thomas, *Leviathan: with selected variants from the Latin edition of 1668*, ed. by Edwin Curley (Indianapolis: Hackett Publishing Company, Inc., 1994).

Hodgson, D. H., *Consequences of Utilitarianism: A Study in Normative Ethics and Legal Theory* (Oxford: Clarendon Press, 1967).

Hollis, Martin, *Trust Within Reason* (Cambridge: Cambridge University Press, 1998).

Honderich, Ted (ed.), *The Oxford Companion to Philosophy* (Oxford: Oxford University Press, 1995).

Hooker, Brad (1990), “Rule-Consequentialism,” *Mind* vol. 99, no. 393 (January 1990), pp. 67–77.

——— (1994a), “Is Rule-Consequentialism a Rubber Duck?” *Analysis* vol. 54, no. 2 (April 1994), pp. 92–97.

——— (1994b), “Compromising with Convention,” *American Philosophical Quarterly* vol. 31, no. 3 (October 1994), pp. 311–317.

——— (1995), “Rule-Consequentialism, Incoherence, Fairness,” *Proceedings of the Aristotelian Society* new series vol. 95 (1995), pp. 19–35.

——— (1996), “Ross-style Pluralism versus Rule-consequentialism,” *Mind* vol. 105, no. 420 (October 1996), pp. 531–552.

——— (2000), *Ideal Code, Real World: A Rule-consequentialist Theory of Morality* (Oxford: Clarendon Press, 2000).

Hospers, John, “Rule-Egoism,” *The Personalist* vol. 54, no. 4 (Autumn 1973), pp. 391–395.

Hume, David (1777), *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (Oxford: Clarendon Press, 1975; based on the 1777 edition).

——— (1779), *Dialogues Concerning Natural Religion and the Posthumous Essays Of the Immortality of the Soul and Of Suicide*, ed. by Richard Popkin (Indianapolis, Ind.: Hackett Publishing Company, 1980; first published in 1779).

Kagan, Shelly (1989), *The Limits of Morality* (Oxford: Clarendon Press, 1989).

——— (1992), “The Structure of Normative Ethics,” *Philosophical Perspectives* vol. 6 (Ethics, 1992), pp. 223–242.

Kalin, Jesse, “Two Kinds of Moral Reasoning: Ethical Egoism as a Moral Theory,” *Canadian Journal of Philosophy* vol. 5, no. 3 (November 1975), pp. 323–356.

Kant, Immanuel, *Toward Perpetual Peace*, in Immanuel Kant, *Practical Philosophy*, trans. and ed. by Mary J. Gregor (Cambridge: Cambridge University Press, 1996).

Kavka, Gregory S. (1978), “Some Paradoxes of Deterrence,” *The Journal of Philosophy* vol. 75, no. 6 (June 1978), pp. 285–302.

——— (1983), “The Toxin Puzzle,” *Analysis* vol. 43, no. 1 (January 1983), pp. 33–36.

——— (1986), *Hobbesian Moral and Political Theory* (Princeton: Princeton University Press, 1986).

Kenny, Anthony, *The Logic of Deterrence* (London: Firethorn Press, 1983).

Korsgaard, Christine (1986), “Skepticism about Practical Reason,” in Christine M. Korsgaard, *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), pp. 311–334; originally published in *The Journal of Philosophy* vol. 83, no. 1 (January 1986), pp. 5–25.

——— (1997), “The Normativity of Instrumental Reason,” in Cullity and Gaut, pp. 215–254.

Kupperman, Joel J. (1980), “Vulgar Consequentialism,” *Mind* vol. 89, no. 355 (July 1980), pp. 321–337.

——— (1981), “A Case for Consequentialism,” *American Philosophical Quarterly* vol. 18, no. 4 (October 1981), pp. 305–313.

Kydland, Finn E., and Edward C. Prescott, “Rules Rather than Discretion: The Inconsistency of Optimal Plans,” *Journal of Political Economy* vol. 85, no. 3 (June 1977), pp. 473–491.

Langenfus, William L., “Implications of a Self-Effacing Consequentialism,” *The Southern Journal of Philosophy* vol. 27, no. 4 (Winter 1989), pp. 479–493.

Lewis, David, “Utilitarianism and Truthfulness,” *Australasian Journal of Philosophy* vol. 50, no. 1 (May 1972), pp. 17–19.

Lewis, H. D. (ed.), *Contemporary British Philosophy: Personal Statements*, fourth series (London: George Allen & Unwin Ltd, 1976).

Lyons, David (1965), *Forms and Limits of Utilitarianism* (Oxford: Clarendon Press, 1965).

——— (1980), “Utility as a Possible Ground for Rights,” *Noûs* vol. 14, no. 1 (March 1980), pp. 17–28.

Mackie, J. L. (1973), “The Disutility of Act-Utilitarianism,” *The Philosophical Quarterly* vol. 23, no. 93 (October 1973), pp. 289–300.

——— (1977), *Ethics: Inventing Right and Wrong* (London: Penguin Books, 1977).

Mankiw, N. Gregory, *Macroeconomics*, second ed. (New York: Worth Publishers, 1994).

McClennen, Edward F. (1988), “Constrained Maximization and Resolute Choice,” *Social Philosophy & Policy* vol. 5, no. 2 (Spring 1988), pp. 95–118.

——— (1990), *Rationality and Dynamic Choice: Foundational Explorations* (Cambridge: Cambridge University Press, 1990).

——— (1997), “Pragmatic Rationality and Rules,” *Philosophy & Public Affairs* vol. 26, no. 3 (Summer 1997), pp. 210–258.

McPhee, *The Control of Nature* (New York: Farrar Straus Giroux, 1989).

Mendola, Joseph, “Parfit on Directly Collectively Self-Defeating Theories,” *Philosophical Studies* vol. 50, no. 1 (July 1986), pp. 153–166.

Mill, John Stuart (1843), *A System of Logic Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*, books IV–VI and appendices (volume 8 of *Collected Works of John Stuart Mill*), ed. by J. M. Robson (Toronto: University of Toronto Press, 1974).

——— (1859), *On Liberty*, in John Stuart Mill, *Essays on Politics and Society* (volume 18 of *Collected Works* on John Stuart Mill), ed. by J. M. Robson (Toronto: University of Toronto Press, 1977), pp. 213–310.

——— (1861), *Utilitarianism*, in John Stuart Mill, *Essays on Ethics, Religion and Society* (volume 10 of *Collected Works of John Stuart Mill*), ed. by J. M. Robson (Toronto: University of Toronto Press, 1969), pp. 203–259.

——— (1873), *Autobiography*, in John Stuart Mill, *Autobiography and Literary Essays* (volume 1 of *Collected Works of John Stuart Mill*), ed. by John M. Robson and Jack Stillinger (Toronto: University of Toronto Press, 1981), pp. 1–290.

Mintoff, Joseph, "Is the Self-Interest Theory Self-Defeating?" *Dialogue* vol. 35, no. 1 (Winter 1996), pp. 35–52.

Monmonier, Mark, *Drawing the Line: Tales of Maps and Cartocontroversy* (New York: Henry Holt and Company, 1995).

Moore, G. E., *Principia Ethica*, revised ed., edited by Thomas Baldwin (Cambridge: Cambridge University Press, 1993).

Moore, Stanley, "Hobbes on Obligation, Moral and Political; Part One: Moral Obligation," *Journal of the History of Philosophy* vol. 9, no. 1 (January 1971), pp. 43–62.

Morgan, Vance G., "Using Group Projects in Business Ethics Courses," *APA Newsletter on Teaching Philosophy*, vol. 95, no. 1 (Fall 1995), pp. 104–106.

Moser, Paul K., "Consequentialism and Self-Defeat," *The Philosophical Quarterly* vol. 41, no. 162 (January 1991), pp. 82–85.

Mulholland, Leslie, "Egoism and Morality," *The Journal of Philosophy* vol. 86, no. 10 (October 1989), pp. 542–550.

Narveson, Jan (1967), *Morality and Utility* (Baltimore, Md.: Johns Hopkins Press, 1967).

——— (1976), "Utilitarianism, Group Action, and Coordination or, Must the Utilitarian be a Buridan's Ass?" *Noûs* vol. 10, no. 2 (May 1976), pp. 173–94.

Nietzsche, Friedrich, *On the Genealogy of Morals*, trans. by Walter Kaufmann and R. J. Hollingdale, ed. by Walter Kaufmann (New York: Vintage Books, 1967).

Nozick, Robert, *Anarchy, State, and Utopia* (New York: Basic Books, 1974).

Parfit, Derek (1978), "Innumerate Ethics," *Philosophy & Public Affairs* vol. 7, no. 4 (Summer 1978), pp. 285–301.

——— (1984), *Reasons and Persons* (Oxford: Clarendon Press, 1984).

Pettit, Philip, "Consequentialism," in Singer 1991, pp. 230–240.

Quinton, Anthony, *Utilitarian Ethics*, second ed. (La Salle, Illinois: Open Court, 1989).

Railton, Peter, “Alienation, Consequentialism, and the Demands of Morality,” *Philosophy & Public Affairs* vol. 13, no. 2 (Spring 1984), pp. 134–171.

Rawls, John (1951), “Outline of a Decision Procedure for Ethics,” in Rawls 1999a, pp. 1–19; originally published in *The Philosophical Review* vol. 60, no. 2 (April 1951), pp. 177–97.

——— (1980), “Kantian Constructivism in Moral Theory,” in Rawls 1999a, pp. 303–358; originally published in *The Journal of Philosophy* vol. 77, no. 9 (September 1980), pp. 515–572.

——— (1999a), *Collected Papers*, ed. by Samuel Freeman (Cambridge: Harvard University Press, 1999).

——— (1999b), *A Theory of Justice*, rev. ed. (Cambridge: Belknap Press, 1999).

Regan, Donald, *Utilitarianism and Co-operation* (Oxford: Clarendon Press, 1980).

Rescher, Nicholas (1975), *Unselfishness: The Role of the Vicarious Affects in Moral Philosophy and Social Theory* (Pittsburgh, Pa.: University of Pittsburgh Press, 1975).

Rescher, Nicholas (1993), *A Study of Pragmatic Idealism, vol. 2: The Validity of Values: A Normative Theory of Evaluative Rationality* (Princeton: Princeton University Press, 1993).

Romer, David, *Advanced Macroeconomics* (New York: McGraw-Hill, 1996).

Ross, W. D., *The Right and the Good* (Oxford: Clarendon Press, 1930).

Sanders, Steven M. (1976), “A Credible Form of Egoism?” *The Personalist* vol. 57, no. 3 (Summer 1976), pp. 272–278.

——— (1978), “Egoism, Self, and Others,” *The Personalist* vol. 59, no. 3 (July 1978), pp. 295–303.

——— (1979), “Egoism Agonistes: A Reply to Berg,” *The Personalist* vol. 60, no. 4 (October 1979), pp. 448–450.

——— (1988), “Is Egoism Morally Defensible?” *Philosophia* vol. 18, no. 2–3 (July 1988), pp. 191–209.

Scanlon, T. M., “Contractualism and Utilitarianism,” in Sen and Williams, pp. 103–128.

Scarre, Geoffrey, *Utilitarianism* (London: Routledge, 1996).

Scheffler, Samuel, *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions* (Oxford: Clarendon Press, 1982).

Schneewind, J. B. (ed.), *Reason, Ethics, and Society: Themes from Kurt Baier, with His Responses* (Chicago: Open Court Publishing Company, 1996).

Sen, Amartya, “Utilitarianism and Welfarism,” *The Journal of Philosophy* vol. 76, no. 9 (September 1979), pp. 463–489.

Sen, Amartya, and Bernard Williams, *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982).

Setiya, Kieran, “Parfit on Direct Self-Defeat,” *The Philosophical Quarterly* vol. 49, no. 195 (April 1999), pp. 239–242.

Shaw, William H. (1980), “Intuition and Moral Philosophy,” *American Philosophical Quarterly* vol. 17, no. 2 (April 1980), pp. 127–134.

——— (1999), *Contemporary Ethics: Taking Account of Utilitarianism* (Oxford: Blackwell Publishers, 1999).

Sidgwick, Henry, *The Methods of Ethics*, seventh ed. (Indianapolis: Hackett, 1981).

Singer, Peter (1972), “Is Act-Utilitarianism Self-Defeating?” *Philosophical Review* vol. 81, no. 1 (January 1972), pp. 94–104.

——— (1974), “Sidgwick and Reflective Equilibrium,” *The Monist* vol. 58, no. 3 (July 1974), pp. 490–517.

——— (1991) (ed.), *A Companion to Ethics* (Oxford: Blackwell Publishers, 1991).

Singleton, Jane, “Moral Theories and Tests of Adequacy,” *The Philosophical Quarterly* vol. 31, no. 122 (January 1981), pp. 31–46.

Slote, Michael (1984), “Satisficing Consequentialism,” part I, *The Aristotelian Society* supplementary volume 58 (The Aristotelian Society, 1984), pp. 139–163.

——— (1985), *Common-Sense Morality and Consequentialism* (London: Routledge & Kegan Paul, 1985).

Smart, J. J. C., “Extreme and Restricted Utilitarianism,” *Philosophical Quarterly* vol. 6, no. 25 (October 1956), pp. 344–354.

Sobel, Jordan Howard, “Rule-Utilitarianism,” *Australasian Journal of Philosophy* vol. 46, no. 2 (August 1968), pp. 146–65.

Sosa, Ernest (ed.), *The Philosophy of Nicholas Rescher: Discussion and Replies* (Dordrecht, Holland: D. Reidel Publishing Co., 1979).

Sprigge, T. L. S. (1965), “A Utilitarian Reply to Dr. McCloskey,” *Inquiry* vol. 8 (1965), pp. 264–291.

——— (1988), *The Rational Foundations of Ethics* (London: Routledge & Kegan Paul, 1988).

Sterba, James P., “From Rationality to Morality,” in James P. Sterba (ed.), *Ethics: The Big Questions* (Oxford: Blackwell Publishers, 1998), pp. 105–116.

Stocker, Michael (1969), “Consequentialism and Its Complexities,” *American Philosophical Quarterly* vol. 6, no. 4 (October 1969), pp. 276–289.

——— (1976), “The Schizophrenia of Modern Ethical Theories,” *The Journal of Philosophy* vol. 73, no. 14 (August 12, 1976), pp. 453–466.

——— (1982), “Responsibility Especially for Beliefs,” *Mind* new series vol. 91, no. 363 (July 1982), pp. 398–417.

Strotz, R. H., “Myopia and Inconsistency in Dynamic Utility Maximization,” *The Review of Economic Studies* vol. 23, no. 3 (1955–1956), pp. 165–180.

Sumner, L. W., *Welfare, Happiness, and Ethics* (Oxford: Clarendon Press, 1996).

Velleman, J. David, “Deciding How to Decide,” in J. David Velleman, *The Possibility of Practical Reason* (Oxford: Clarendon Press, 2000), pp. 221–243; originally published in Cullity and Gaut, pp. 29–52.

Williams, Bernard (1972), *Morality: An Introduction to Ethics* (Cambridge: Cambridge University Press, 1972).

——— (1973), “Deciding to Believe,” in Bernard Williams, *Problems of the Self* (Cambridge: Cambridge University Press, 1973), pp. 136–151.

——— (1981), “Preface,” in Bernard Williams, *Moral Luck: Philosophical Papers, 1973–1980* (Cambridge: Cambridge University Press, 1981), pp. ix–xi.

——— (1982), “The Point of View of the Universe: Sidgwick and the Ambitions of Ethics,” in Bernard Williams, *Making Sense of Humanity and Other Philosophical Papers, 1982–1993* (Cambridge: Cambridge University Press, 1995), pp. 153–171.

——— (1985), *Ethics and the Limits of Philosophy* (Cambridge: Harvard University Press, 1985).

Winters, Barbara, “Believing at Will,” *The Journal of Philosophy* vol. 75, no. 5 (May 1979), pp. 243–256.

Woodward, Bob, *Maestro: Greenspan’s Fed and the American Boom* (New York: Simon & Schuster, 2000).