INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700  800/521-0600
Seeking the standpoint of rationality

Tierney, Robert Barkley, M.A.

Rice University, 1993
RICE UNIVERSITY

SEEKING THE STANDPOINT OF RATIONALITY

by

ROBERT B. TIERNEY

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
MASTER OF ARTS

APPROVED, THESIS COMMITTEE

Larry Temkin, Director
Professor of Philosophy

Richard Grandy
Professor of Philosophy

Peter Mieszkowski
Professor of Economics

Houston, Texas
April, 1993
ABSTRACT

Seeking the Standpoint of Rationality

by

Robert Tierney

Theories of practical rationality which are based upon the concept of the maximization of preference satisfaction are open to criticism on the grounds that they alienate the agent from his actions, and from the source of his actions in his moral emotions. This general analysis of rational motivation can be further supported by a close analysis of the rational strategy known as "precommitment". Ordinarily taken to exemplify an ideal strategy of rational maximization, upon closer inspection the rational motivation involved in precommitment is typically not maximization of preference satisfaction, nor should it be. These arguments show that the agent would be better off in terms of preference maximization if he believed some other theory of practical rationality. If the preference satisfaction model recommends some other theory of practical rationality, then it no longer directly provides the agent with reasons for action, in which case it no longer serves as a theory of practical rationality at all.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Larry Temkin, for his many helpful and illuminating comments on this project, and for showing patience in allowing this project to evolve through the early stages in which it was none too clear what I was up to. I would also like to thank Peter Miezskowski for the long, and sometimes troubling, conversations which forced me to clarify and refine my ideas, and for helping me to better understand the perspective of economics. Thanks are also due to Richard Grandy for his incisive and instructive comments on this thesis.

I would also like to thank my fellow graduate students for offering their comments on, and criticisms of, this project. Greg Ealick and David Schwartz were especially generous with their time and insight, and so helped to make this thesis better than it otherwise would have been.

I am grateful, too, for the insight, encouragement and sympathy that Kelly Cowling shared with me in the early stages of this project, though, sadly, we were not able to see it through to its completion together.

I must give a special thanks to my good friend Bob Bifulco for obstinately pushing me to be clear in my thoughts, and for forcing me to argue for things that I had wanted to
assume. Our many conversations over the years have done much to clarify in my mind just what is at stake in the broader context in which these argument occur, and in so doing have helped me to regain a sense of purpose in my work at times when it seemed to be slipping away.

I am truly grateful to all who have supported me in this endeavor.
TABLE OF CONTENTS

INTRODUCTION 1

CHAPTER 1
I: Williams' argument against Utilitarianism 5
II: A parallel argument against Self-Interest Theory 23

CHAPTER 2
Introduction 41
I: Time Preferences, Weakness of the Will and Precommitment 42
II: Precommitment and the Motive to Maximize 53

CHAPTER 3
Introduction 71
I: Whether Self-Interest Theory is Self-Defeating 74
II: Whether Self-Interest Theory is Indirectly Self-Defeating and Not Directly Self-Defeating 80
III: Whether Self-Interest Theory is Self-Effacing and Not Self-Defeating 84

Introduction 84
A. The Argument That Self-Interest Theory is Self-Effacing and Not Self-Defeating 87
B. Self-Interest Theory Cannot be Completely Self-Effacing and Still Be a Theory of Practical Rationality 89
C. The Regress Argument Against Self-Effacing Theories Practical Rationality 96
D. The Empirical Argument That Self-Interest Theory is Self-Defeating Even If It Is Self-Effacing 102

E. Problems Grounding Self-Interest Theory If It Is Self-Effacing 109

CONCLUSION 116

Bibliography of Sources Cited 120
INTRODUCTION

Self-Interest Theory is a theory of practical rationality. That is, it is a theory about what it is most rational for an agent to do. It claims that what it is most rational for an agent to do is whatever will make his life go as well as possible. In this thesis I argue that Self-Interest Theory is not the true or the best theory of rationality.

There are three versions of Self-Interest Theory. The first version claims that an agent makes his life go as well as possible if he maximizes the satisfaction of his preferences over the entire course of his life. My arguments are primarily focused on this version of the theory. Another version of Self-Interest Theory claims that an agent's life goes as well as possible if he maximizes his happiness over the course of his entire life. Happiness is construed in this context as some sort of mental state. To an extent to be clarified later, my arguments apply to this theory as well. A third version of Self-Interest Theory claims that an agent's life goes as well as possible if he maximizes over some objective list of goods. I do not be address this last version of the theory at all in my arguments.

I make several types of arguments against Self-Interest

---

1 Here I am following Parfit. See Parfit 1984, pp. 3-4.
Theory. In Chapter 1, I argue that when consistently thought through, Self-Interest Theory appears to provide us with a conception of agents and how they ought to live their lives which is unacceptable. Adopting Self-Interest Theory, I argue, would mean viewing one's life externally, so to speak, in a way that is incompatible with some of our deepest intuitions about how lives ought to be lived.

The argument in Chapter 2 supports the argument in Chapter 1 by examining a particular case of what is normally taken to be a paradigmatic instance of self-interested maximization and showing that the agent would not be motivated by maximization in such a case. I examine a case in which an agent uses a strategy for overcoming weakness of the will known as "precommitment" in order to see how best to interpret his motives. The best interpretation of the agent's motives, I argue, is that they are something other than utility maximization. The examination of this particular case lends support to the more general argument of the first chapter that Self-Interest Theory is not the best interpretation of ideal human practical rationality.

The third chapter has a more complicated structure. I first argue that Self-Interest theory is self-defeating based upon the phenomenological evidence of the first two chapters. I then consider the response that Self-Interest Theory is only
indirectly self-defeating and not damagingly self-defeating and find it wanting. I then consider a second line of defense for Self-Interest Theory: that it is what is called "self-effacing" and not self-defeating. That is, Self-Interest Theory might best guide our actions in telling us to believe some other theory of rationality.

If Self-Interest Theory can be self-effacing and not self-defeating, it may be able to escape not only the force of the criticism that it is self-defeating, but also the criticisms against it made in the first two chapters. For it would tell us to believe some theory which did not conflict with our intuitions in the ways described in the first two arguments. I make four arguments which apply to Self-Interest Theory if it is a self-effacing theory. The first of these is that a fully self-effacing theory could not give an agent reasons for action, and so would not be a theory of rationality at all. The second is a reductio ad absurdum of the idea of a self-effacing theory of rationality. The third argument is an empirical argument that Self-Interest Theory would be self-defeating even if it was self-effacing. Lastly, I argue that one important method of grounding Self-Interest Theory would be lost if it were self-effacing, namely, reflective equilibrium.

I do not take these final four arguments to be
conclusive. Each requires some rather significant assumptions. To provide more powerful evidence these arguments would require further theoretical development in such areas as theory of practical rationality, theory of theoretical rationality and epistemology. This being the case, the force of the arguments of the first two chapters may also be weakened. However, my arguments provide good ground for believing that we should not accept Self-Interest Theory at the primary level as a theory of practical rationality which guides our lives and give us reason to doubt that it survives as a tenable theory at the meta-level.
CHAPTER 1

I. Williams' Argument Against Utilitarianism

Two of Williams' chief criticisms of Utilitarianism are: (1) it alienates one from one's moral emotions, and (2) it alienates one from one's actions. In this section of Chapter 1 I will explicate Williams' arguments in support of these two criticisms. In the next section I will argue that parallel criticisms apply to Self-Interest Theory. Thus, if Williams' arguments against Utilitarianism succeed, then similar arguments against Self-Interest Theory will succeed. My aim, then, is to show that Self-Interest Theory, when adopted by an agent, alienates him from his goals and values, as well as from his actions.

What Utilitarianism instructs us to do is, roughly, to act so as to promote the greatest good for the greatest number. There are obvious questions about how we are to cash out the term "number" in this formulation. For instance, there is the question of whether utility ought to be maximized for all sentient beings, or simply maximized for all persons. There is also the question of whether utility ought to be maximized for all future persons, and the various problems
surrounding this issue. I will not discuss these issues within this thesis, but will simply assume that Utilitarianism is concerned with maximizing utility for all and only persons, and that the discussion can get underway without first sorting out these issues.

Assuming that the notion of maximization is clear enough, the remaining term from the description of Utilitarianism which is given above which requires clarification is "good". The term "good" in Utilitarian Theory has been used primarily to designate either "preference satisfaction" or "happiness". In the context of discussions of Utilitarianism, happiness is usually considered to be a mental state or sensation. In the case of either definition of utility, however, the notion of what the good is for persons remains largely unspecified.

To see the force of this latter claim, let us first consider the version of Utilitarianism which construes "good" as "satisfaction of preferences". On this version of the theory, maximization of preference satisfaction presupposes that agents already have preferences to satisfy. These preferences range from what Williams calls "our deepest projects and commitments" to our momentary whims. Utilitarianism does not speak to the content of these

---

2 See Parfit, 1984, Part IV.

preferences, neither with regard to what they are nor to what they ought to be, but only urges that we satisfy them to the greatest degree possible. In this respect, Utilitarianism presupposes that agents already value certain things—at least in the minimal sense that preferring a thing is regarded as valuing it—and that these preferences provide the content over which the ideal Utilitarian agent is to maximize.

If, on the other hand, the Utilitarian project is regarded not as the maximization of the satisfaction of preferences but, rather, as the maximization of happiness, then its conception of the good still remains largely unspecified. Consider the fact that agents, generally speaking, find the source of their happiness in performing certain particular activities or enjoying various particular objects of desire. The ends sought here might range from fighting for a political cause to enjoying a sunset. The point is, that a large measure of the happiness that agents enjoy is gained through certain particular projects and activities. Utilitarianism does not speak to what these projects ought to be, or as a matter of fact are likely to be, but only urges that we ought to promote people's ability to engage in them so as to maximize happiness. Thus, whether the Utilitarian relies upon the concept of happiness or of preference satisfaction in cashing out his conception of the good, that concept of the good will be largely without
particular content.

In describing Utilitarianism I have used the word "projects". This is a term used by Williams to indicate those endeavors and goals in the lives of individuals which shape those lives in various ways. These projects may include seeking the bare necessities of life, as well as pursuing objects of taste. These pursuits may also include cultural, creative, or intellectual endeavors which, while they may not be considered necessities for physically sustaining life, may nonetheless involve a deep attachment on the part of the individual. Williams writes that "...an individual person has a set of desires, concerns or, as I shall call them, projects, which help to constitute a character".  

One will tend to be more deeply involved and identify more closely with certain of these projects. Williams refers to these projects to which one is more deeply attached as "commitments". In using these terms Williams is not endorsing certain pursuits as opposed to others. It is merely that by using the language of projects and commitments Williams is stressing the active and engaged aspect of

---

5 Williams, 1981 (a), p. 5.
6 Williams, 1973, p. 112.
individual lives. For it is all too easy to forget when discussing, in theoretical terms, the agents who are the recipients of utility that they are not merely hedon receptacles or passive utility consumers. Thus, in speaking in terms of "projects" and "commitments" Williams is using a language which attempts to preserve the sense of agents as active, involved persons, beings with particular pursuits and ideals engaged in the world, without surreptitiously adding content to the notion of good under discussion.

We might think of "preference satisfaction" as referring, at least in part, to the successful performance of projects. Our preferences would be whatever projects, however great or small, that we choose to engage in. I will not defend this interpretation of Utilitarian Theory. I will merely carry out the arguments of this chapter and the next with the understanding that the talk about the success of projects is directly applicable to notions of preference satisfaction.

Alternatively, while it may not be the case that an act which most promotes the successful completion of projects always maximizes happiness, I believe that under ordinary circumstances of human life the maximal success of our projects and our maximal happiness will tend to be produced by the same acts. To the extent that the reader takes them to

---

7 For criticism of this habit of thought see Korsgaard.
diverge, he may make an argument against the applicability of these criticisms to versions of Utilitarianism which employ a mental state theory of the good. For it could be argued that it is the focus on the maximal success of projects, rather than the agent's happiness, that leads to the problems involved in the following criticisms of Utilitarianism and Self-Interest Theory. In this thesis I will neither develop such criticisms nor respond to them.\textsuperscript{8}

Reinterpreting the rough definition of Utilitarianism given at the beginning of this chapter in Williams terms, we can say that the object of Utilitarianism is to maximize the satisfaction of projects. It follows from this that Utilitarianism, itself, is a second-order project that seeks to maximize the outcome of first-order projects. Thus, the aim of the Utilitarian project, the "good" which it seeks to promote, gains its content only indirectly, from the first-order projects. It could be the case that one of an agent's basic first-order projects happened to be the maximization of

\textsuperscript{8} Williams restricts much of his discussion to Utilitarian theories which employ a preference satisfaction theory of the good. Williams writes:

...I shall in general assume, along with most modern writers in philosophy and economics, that in talking of happiness or utility one is talking about people's desires or preferences and their getting what they want or prefer, rather than about some sensation of pleasure or happiness. (Williams, 1973, p. 80)
utility. However, that project would never get off the ground unless at least one person had some other first-order project the successful outcome of which the agent could promote. Williams writes: "Unless there were first-order projects, the general utilitarian project would have nothing to work on, would be vacuous".⁹ Thus, Utilitarianism requires some first-order projects to serve as terms in its utility calculations if it is to have content.

The successful promotion of the "maximally harmonious realization of those projects" demands that there be no bias which gives preferential treatment to any project or series of projects.¹⁰ For if the content of any one project, or collection of projects, made some special claim to having priority in the utility calculus it would introduce the possibility that the aims of the project would be promoted at the expense of overall utility maximization. Therefore, the aims of a given project must not be given consideration in the utility calculus except formally, that is, in terms of whether it tends to promote or retard the maximally harmonious realization of projects. The standpoint of Utilitarian assessment, then, must be from outside of every first-order project.

⁹ Williams, 1973, p. 110.

¹⁰ This phrase is taken from Williams, 1973, p. 110.
Another feature of the Utilitarian standpoint is that it has no given concern with how a given state of affairs is produced or who produces it except as it relates to the sum total of utility. Utilitarianism is simply concerned with producing one particular state of affairs—that in which utility is maximized. As far as the project of utility maximization is concerned, there can be no significance to the fact that Smith performs a given act rather than that Jones performs a similar act unless, and only to the extent that, that change in persons performing the act entails a further change in the amount of utility produced. Thus, if the world is causally structured such that equal amounts of utility are produced whether Smith performs some act A or Jones performs some act B, then from the Utilitarian standpoint we ought to be indifferent between them. However, should either A or B promote a slightly higher degree of utility, then that act ought to be done.

Just as the mere fact that Smith rather than Jones performs a given act considered irrespective of causal consequences is of no significance from the Utilitarian standpoint, neither can the mere fact that I, rather than Smith, performed that act have any import from the Utilitarian standpoint. As Williams writes, from the Utilitarian standpoint "[t]here is no comprehensible difference which consists just in my bringing about a certain outcome rather
than someone else's producing it."\textsuperscript{11} For "[g]ranted that states of affairs have been adequately described in causally and evaluatively relevant terms, it makes no further comprehensible difference who produces them".\textsuperscript{12} The mere fact that I, rather than someone else, performed a given act is of no concern whatsoever from the Utilitarian standpoint.

Certainly in some circumstances one might have to consider the truth of general claims about how the Utilitarian calculus would be affected if one person rather than another performed a given act. One such claim would be, for example, the psychological claim that thwarting one's own project is a particularly bitter experience. Similarly, one might have to consider general causal claims such as the claim that one's ability to control the outcome of his own project is greater than his ability to control the outcomes of the projects of others.

Such claims as these can be assessed and applied without reference to the first person standpoint. For these are issues of whether a given agent described as having certain objective properties or characteristics will produce a given outcome. The entire situation relating to the production of the state of affairs can be described from outside of the

\textsuperscript{11} Williams, 1973, p. 96.

\textsuperscript{12} Williams, 1973, p. 95.
first person standpoint. The fact that certain psychological claims relating to the frustration of an agent at having to thwart his own project applies to me in a given case, rather than to Smith, is of no significance whatsoever from a Utilitarian standpoint. Put succinctly, "'it's me' can never in itself be a morally comprehensible reason".\textsuperscript{13} Williams writes of the fact that Utilitarian theory seeks to get beyond the first person standpoint that:

\begin{quote}
It helps to explain why consequentialism [of which Utilitarianism is a species] can seem to express a more serious attitude than non-consequentialist views, why part of its appeal is to a certain kind of high-mindedness. Indeed, that is part of what is wrong with it.\textsuperscript{14}
\end{quote}

The next question, then, is why Williams takes the impersonal standpoint to be a problematic feature of Utilitarianism.

Thus far we have discussed two main features of Utilitarianism: (1) that Utilitarianism is a second-order project, and (2) that the standpoint of Utilitarian assessment is the impersonal standpoint. In what follows, I will attempt to explain just what Williams' criticisms of these features of Utilitarianism are. This can best be done by making use of a specific example, and none better than Williams' own.

The gist of the example is that there is a man, call him

\begin{itemize}
\item[\textsuperscript{13}] Williams, 1973, p. 96.
\item[\textsuperscript{14}] Williams, 1973, p. 96.
\end{itemize}
Jim, who is in South America on a botanical expedition. Jim accidently wanders into a South American village where a death squad is about to execute some of the Indian villagers in reprisal for recent political protests. Ten Indians have been selected at random and are tied against the wall. The captain of the squad was just about to give his man, Pedro, the order to fire when Jim stumbled onto the scene. After questioning Jim about his identity and business in the area, the captain tells Jim that as an honored guest from another country Jim may have the privilege of killing one of the Indians himself. In honor of the event, the other nine Indians will be set free. If Jim refuses, all ten Indians will be shot by Pedro. The Indians against the wall know full well what the situation is, and are begging Jim to shoot. Jim considers trying to overpower the gunmen, but he knows that he cannot.\footnote{Williams, 1973, pp. 98-99.}

Williams tells us that he provides us with this case for three reasons: (1) to consider whether Utilitarianism provides the right answer, (2) to consider whether the answer that Utilitarianism must suggest—that Jim ought to take one life in order to save nine—can be obviously right in the way that Utilitarianism suggests that it is, and (3) to gain an understanding of what kind of considerations go into finding an answer to this kind of moral dilemma.\footnote{Williams, 1973, p. 99.}
really takes strong issue with what he takes to be Utilitarianism's answer to this problem (1). Rather, he is more concerned with how Utilitarianism arrives at its answer (3), and how this relates to the uncomfortable conclusion that this is the obviously right answer (2).

Williams states early on in the essay that the aim of the examples that he will later provide, and this includes the case of Jim, is not just to elicit moral intuitions which will either support or fail to support our theory. Instead, "the aim is to lead to reflections which might show up in greater depth what would be involved in living with these ideas".17 He is writing here about what it would mean to live with the moral conceptions involved in Utilitarianism. The issue is whether or not, once we see how Utilitarianism requires us to view the world, we would be willing, or even capable, of adopting such a view, and whether one would be capable of sustaining a coherent moral life were he to do so. We must now turn to the Utilitarian analysis of this case.

When Jim faces the choice of whether or not to shoot the Indian, one of his considerations is likely to be that in one case he would be killing a human being, and in the other he would be killing no one. Jim is likely to have a very strong moral sense that it is wrong for him to kill an innocent

17 Williams, 1973, p. 78.
person. We might suppose that he has a strong commitment to preserving rather than taking human life.

One thing is clear, it is not the case that either Jim shoots one Indian or Jim leaves Pedro no choice but to shoot ten Indians. For Pedro is as much a person as Jim, and can make a choice in this situation just as Jim can. Nevertheless, it may be true that Pedro will undoubtedly shoot the ten unless Jim shoots the one.

Despite these considerations of agency, for the Utilitarian the analysis of this case is straightforward. From the Utilitarian standpoint Jim quite simply has a choice: ten Indians killed or one Indian killed. This is true because as a Utilitarian one is only interested in that state of affairs which represents the maximization of utility.

As regards how this state of affairs was produced, it must be remembered that for Utilitarianism all causal connections are on a par. Therefore, "it makes no difference, so far as that goes, whether the causation of a given state of affairs lies through another agent, or not".18 "For Pedro's killing the Indians to be an outcome of Jim's refusal, it only has to be causally true that if Jim had refused, Pedro would

18 Williams, 1973, p. 94.
have done it". It follows that "[i]f I am ever responsible for anything, I am just as responsible for things that I allow or fail to prevent, as I am for things that I myself, in the more everyday restricted sense, bring about". Williams calls this "negative responsibility". Jim cannot claim that he is not responsible for the lives of the Indians.

Jim cannot stand on some deontological prohibition against killing. For, as I have said, the mere fact that one person, rather than another, produced a given state of affairs can have no significance from the Utilitarian standpoint, even if one of the persons is me. "Granted that the states of affairs have been described in causally and evaluatively relevant terms, it makes no further comprehensible difference who produces them." It is true that there may be some disutility involved in Jim having to override strong moral convictions in order to shoot the Indian. However, the unpleasantness for Jim caused by this action is likely to be trivial compared with the value of the lives that are at stake. From the Utilitarian standpoint the proper choice for Jim remains clear. Were Jim

---

20 Williams, 1973, p. 95.
21 Williams, 1973, p. 95.
22 Williams, 1973, p. 95.
to refuse to shoot the Indian he might be accused of squeamishness, of being morally self-indulgent, in that he is more interested in preserving his own innocence than in preserving the lives of the Indians.

The squeamishness appeal exemplifies an attitude toward human action commensurate with Utilitarianism. For it asks us to consider our deepest moral convictions as mere feelings which can readily be added into the utility calculus like any other factor. This is the distinctly Utilitarian way of regarding these feelings. If one consents to so regarding one's moral feelings and commitments, then the answer to the dilemma is a foregone conclusion. For even some rather severely unpleasant feelings matter relatively little when compared with the value of nine innocent lives. Utilitarianism asks us to consider these convictions against taking life as merely some feelings amongst others, rather than as "emotional expressions of the thought that to kill is wrong". 23

Utilitarianism demands that we view all first-order projects, including our own deepest projects and commitments, such as the commitment not to take an innocent human life, from a place outside of all such projects. They must be regarded merely as one set of projects amongst many such sets.

23 Williams, 1973, p. 103.
Williams writes of the appeal to do so, and of our resistance to such a move that:

The reason why the squeamishness appeal can be very unsettling, and one can be unnerved by the suggestion of self-indulgence in going against utilitarian considerations, is not that we are utilitarians who are uncertain what utilitarian value to attach to our moral feelings, but that we are partially at least not utilitarians, and cannot regard our moral feelings merely as objects of utilitarian value. Because our moral relation to the world is partially given by such feelings, and by a sense of what we cannot 'live with', to come to regard those feelings from a purely utilitarian point of view, that is to say, as happenings outside one's moral self, is to lose a sense of one's moral identity; to lose, in the most literal way, one's integrity. At this point utilitarianism alienates one from one's moral feelings...24

Once one is asked to live a life where one's standpoint of moral assessment lies beyond one's deepest projects and commitments, the sense, the very point, of one's moral life becomes jeopardized. Those convictions and intuitions which made one's life intelligible become mere "happenings outside one's moral self". One has adopted a standpoint which alienates one from the convictions which give one's life meaning and purpose.

In the example about Jim, it was not that Jim controlled Pedro and therefore had responsibility for the lives of the Indians; rather, it was the case that the projects of Pedro

24 Williams, 1973, pp. 103-104.
and his captain shaped the projects of Jim. This situation will not be the exceptional case. For one's projects will typically causally interact with the projects of all those around him, such that one will typically have the ability to promote a better or worse outcome of any number of projects engaged in by others. But for each individual, his projects are the projects of only one individual considered against the projects of an indefinite number of other people. One's own projects and commitments enter the utility calculus as just another set of projects and commitments. Therefore, one's "own decisions as a utilitarian agent are a function of all the satisfactions that he can effect from where he is: and this means that the projects of others, to an indeterminately great extent, determine his decision".25 Thus, the weight of the utility calculus could at any moment overrule the dictates of our deepest projects and commitments; and the utility calculus has a prior claim on our actions. Thus, given the projects of the ten Indians, of Pedro, and of Pedro's captain, Jim is obliged to make it a part of his projects to shoot the Indian.

If this is right, there is a curious sense in which, from the moral (Utilitarian) standpoint we do not really own—that is we have no indissoluble claim upon—any of our projects or commitments, or the actions which would flow from them. We

are only "the agent of the satisfaction system" where that system has the ultimate claim upon our lives. 26 How we are to act at any moment "is a question of what causal levers, so to speak, are at the moment within our reach", and how our own projects and the projects of others have structured the causal scene. 27 Once one has accepted this more or less unlimited demand upon his actions, upon which projects or commitments he may or may not act upon, he has in a serious sense relinquished ownership of his actions. Thus, Williams writes:

The point is that [the agent] is identified with his actions as flowing from projects and attitudes which in some cases he takes seriously at the deepest level, as what life is about (or, in some cases, this section of his life--seriousness is not necessarily the same as persistence). It is absurd to demand of such a man, when the sums come in from the utility network which the projects of others have in part determined, that he should step aside from his own project and decision and acknowledge the decision which the utilitarian calculation requires. It is to alienate him in a real sense from his actions and the source of his action in his own convictions. It is to make him into a channel between input of everyone's projects, including his own, and an output of optimific decision; but this is to neglect the extent to which his actions and his decisions have to be seen as the actions and decisions which flow from the projects and attitudes with which he is most closely identified. 28

Thus, in asking us both to consider ourselves to be, and behave as if we were, "primarily agents of a universal

26 Williams, 1973, p. 115.


28 Williams, 1973, pp. 116-117. Italics are Williams'.
satisfaction system", Utilitarianism alienates us from the projects and commitments which serve to constitute our characters, and so too does it alienate us from the actions which would flow from those projects and commitments.
II. A Parallel Argument Against Self-Interest Theory

In this section I shall make an argument against Self-Interest Theory which parallels Williams' argument against Utilitarianism. The thrust of this section of the thesis is that if Williams' criticism of Utilitarianism succeeds in casting doubt on Utilitarianism as a moral theory, then parallel criticisms ought to cast doubt on Self-Interest Theory as a theory of rationality. There are many possible responses to Williams' criticism of Utilitarianism. I will not take up these critical responses in this chapter, though I will examine one type of response in Chapter 3. One could also argue that other moral theories, or theories of rationality, other than Utilitarianism, also produce a kind of alienation.\(^29\) However, it is not my task in this thesis to investigate these issues. My primary task is not to offer a deeper analysis of Williams or to investigate the issue of alienation in other domains, but to demonstrate some problems with Self-Interest Theory. In the course of the discussion I also will suggest what might be a source of these particular difficulties. Having expressed the limits of this thesis, I now move on to consider how certain issues of alienation arise in the context of Self-Interest Theory.

\(^{29}\) See Williams on Kant in Williams, 1981 (a) and 1981 (b).
In the first part of this chapter I followed Williams in showing how Utilitarianism alienates an agent from his convictions and moral emotions as well as from his actions. I described three distinct features of Utilitarianism: (1) it is a second-order project, (2) it is ultimately concerned with producing a state of affairs—the greatest good for the greatest number, and (3) its concept of good consists of the success of first-order projects. These features of Utilitarianism place certain requirements upon the standpoint of Utilitarian assessment. In that it is a second-order project which aims to promote the maximally harmonious realization of first-order projects, the standpoint of Utilitarian assessment must be outside of any first-order projects. In addition, in that its sole aim is to produce a certain state of affairs, from the standpoint of Utilitarian assessment it is a matter of indifference that certain set of projects are my projects. Utilitarianism demands that I adopt the impersonal standpoint in deciding which action to take. I will now move on to consider whether the standpoint of assessment for Self-Interest Theory shares these two features with the Utilitarian standpoint of assessment.

Put in Williams' terms, the aim of Self-Interest Theory is the maximally harmonious realization of my projects. Like Utilitarianism, Self-Interest Theory: (1) is a second-order project ultimately concerned with promoting a certain state of
affairs—in this case the greatest good for the individual agent, and (2) its concept of the good consists in the satisfaction of first-order projects. It is evident from these features of Self-Interest Theory that the standpoint of assessment for that theory will be from outside of any particular project. In this requirement for its standpoint of assessment, Self-Interest Theory is identical to Utilitarianism. However, Utilitarianism and Self-Interest Theory differ in the second feature of the standpoint of assessment. For Self-Interest Theory, unlike for Utilitarianism, it does matter whether or not a project is mine. Indeed, it is crucial whether or not a project is mine if I am a Self-Interest Theorist. For this is what determines whether I shall give it any weight or not in making my decision about what I ought rationally to do.

My criticism of Self-Interest Theory is based upon the comparison between many persons at one time with one person across time, or, more accurately, the interests of many people at one time with the interests of one person across time. I attempt to show that the temporally neutral standpoint is relevantly similar to the impersonal standpoint, such that they give rise to a similar sort of alienation when these standpoints are construed as being outside of any first-order projects. The basis of the similarity is this: though my future desires may, in an important sense, be mine, they are
not present to me now. In fact, those of my future desires which I do not presently have are no more experientially present to me than the desires of another person. Therefore, if the alienation produced by Utilitarianism has its roots in problems arising from an agent attempting to be obedient to the dictates of a decision procedure which is based upon the maximization of the satisfaction of preferences or aims, only a minority of which does he share in, then this problem could arise equally well in Self-Interest Theory. Self-Interest Theory would be open to essentially the same sort of criticism that Williams makes against Utilitarianism.  

---

30 There is much to be said about this issue. Thomas Nagel argues at length against the view that an agent's future ought to be of concern to him because he, at present, has desires relating to it. See Nagel, 1970, Part II. He writes:

A person's future should be of interest to him not because it is among his present interests, but because it is his future. He already stands in a far stronger and more important relation to his future and the desires he can expect to experience than could possibly be established by any desire which might assail him at present. The latter sort of connection allows him to reach toward something outside himself; the former depends upon an acknowledgement that certain things are not outside to begin with, and that events in his future hold an interest for him now because they belong to a single person of whom the present segment is merely one stage. (Nagel 1970, pp. 42-43. Italics are Nagel's)

I will not respond to Nagel's arguments directly, but in following Williams I will be supporting a position which, in many ways, is contrary to that of Nagel. A little later in this chapter I will argue, following Parfit, that some judgements which Nagel describes as being made from a temporally neutral standpoint can equally well be described as occurring from a temporally relative standpoint, and that this
Much of the argument hangs upon what the relevant notion of "sharing" in desires means. If, at the outset, one simply defines this notion in terms of whether a desire, at some point during the agent's life, will exist for the agent as a present desire, then the argument is over. For, in that case, there would simply be no basis of analogy between the interests of one person across time and the interests of many people at one time. One would plainly and simply always share in, or own, all his past, present and future desires, but never have nominally the same preferences as another agent. Thus, the two cases would remain entirely distinct. However, I will argue that Williams' criticism of Utilitarianism operates not on the basis of whether a preference is nominally one's own or some other's, but on the basis of whether or not the outcome of the decision procedure is based upon preferences that one experiences as real and present. The question, then, is: in what sense is it necessary for an agent to retain a sense of his projects and commitments as his own when adopting a given decision-making procedure such that standpoint is incompatible with Self-Interest theory's standpoint of assessment.

It is possible that the best theory of rationality will require the agent to take a temporally neutral standpoint in certain instances of decision making, such as those concerned with satisfying mere tastes or preferences. This component of the theory might look like Self-Interest Theory. For other sorts of decisions, such as decisions involving values, the agent may be required to take a temporally relative standpoint. In Chapter 2 I will argue that such decisions requiring temporally relative judgements and motives extend beyond what would ordinarily be considered moral decisions.
he does not alienate himself from them in the ways previously discussed.

In order to answer this question, it would be helpful to make use of an example. Suppose that I am a composer of music. What I most want to do now is write a symphony. The intuitive idea for the symphony has been with me for a long time, but it has only recently been clear to me such that I feel that I could execute the idea. I feel inspired to write the symphony. It is a project into which I could put my heart and soul. There is, however, a problem. It will take a great deal of time and energy to complete the symphony. Since I am a terrible perfectionist it will take several years to complete it. Even if I complete the symphony I have good reason to believe that it will be appreciated by only a small audience. I won't be able to make a lot of money by writing it. The time that would be spent writing the symphony is time that could have been spent working on less demanding projects such as writing pop music. I have a real flair for writing pop music. I could gain considerable financial security and material comfort if I devoted myself to writing pop songs. Time spent working on the symphony would mean time not earning money by writing pop tunes or cultivating professional relationships in the pop music business. There is a significant price to be paid if I am to write the symphony.
Though what I most want to do now is compose the symphony, I also know that later in life I am likely to feel some regret if I do so because of the financial opportunities that I would lose. I believe that as I get older my values will become worn away by social pressures. I think of this as a sort of moral degeneration such that I will no longer value creativity, artistic expression, and beauty as highly as I do now. Given that a greater proportion of my life will be taken up by conservative concerns than by my idealistic concerns, the former will tend to slightly outweigh the latter in terms of overall preferences to be satisfied. Therefore, my life will be slightly better overall if I abstain from what I, at present, regard as my deepest and most important project: writing the symphony. It will be sacrificed for concerns that I do not now regard as especially important.

The example serves to highlight the fact that any project that we might have, no matter how deep, can in principle be deemed unworthy by Self-Interest Theory. That is, the cumulative weight of our future preferences could outweigh the utility that any present preference has accrued. This is so even if now I do not have the slightest interest in, or even have a positive dislike for, these future preferences and projects. Thus, the weight of future preferences can always overrule present projects even though I do not, and cannot, experience those aims as present and real.
This is not to say that I will never come to experience those future projects as real, certainly I may. The point is that I only exist in the "now", I can only experience the present. It is a striking thing to realize that each moment of one's life can be at the beck and call of future which as yet experientially unreal. There is a very real sense in which one can only live in the present moment. To be ever called out of the moment by a future which is by its very nature experientially unreal is in an important sense to be called out of life. It is in this sense that the temporally neutral standpoint is akin to the agent-neutral standpoint of Utilitarianism.

The example may be said to be unfair because it relies upon what many persons would take to be a valuable project as the present aim which stood in jeopardy of being overruled. Many strongly felt present aims are not so worthwhile. An intense craving for heroin, or the desire to shoot one's neighbor might be an intensely felt aim that we would hope that persons would not succumb to, if not for moral considerations, then certainly for the sake of their own future satisfaction.

This objection is true as far as it goes, but it misses the point. I have in no way attempted to assess the merit or rationality of any aims, nor to say that present aims ought
always to be acted upon.\textsuperscript{31} I am merely attempting to elucidate a phenomenological fact. We can imagine any number of aims which, were we to sacrifice them, would require a rather severe sort of detachment. One might think here of projects for personal achievement, a deeply romantic relationship, or various moral projects as examples of projects that might be overruled by Self-Interest Theory.

Since the scope of the future is much greater than the scope of the present it contains many more desires than the present. Present aims are just a few among the many aims we will or might come to have over the course of our lives. Just as in the Utilitarian calculations where I am one person

\textsuperscript{31} The best theory might be something like Parfit's Critical Present Aim Theory which allows for rational criticism of an agent's ends, but might avoid the sort of alienation discussed here in that it gives rational and motivational priority to present aims. Critical Present Aim Theory could also allow for a maximization element in practical rationality within a limited realm such as that of tastes and preferences.

In fact, Parfit makes use of examples similar to mine in order to argue for the Present Aim Theory of rationality. See Parfit, 1984, Part II Chapter 6. However, while Parfit uses the examples as a means of attempting to show that certain decisions which run counter to the requirements of Self-Interest Theory seem intuitively rational, I make a different argument. My use of the examples centers around the claim that deciding such cases in terms of Self-Interest Theory would produce a kind of alienation in the agent, and that this alienation should indicate to us that there is something wrong with the view that Self-Interest Theory asks us to take of our lives. Further, despite the fact that I make use of examples similar to those of Parfit, it is not the aim of this paper to argue for Critical Present Aim Theory or any other theory of rationality. My aim is wholly negative: to criticize Self-Interest Theory.
amongst the many, present aims are outweighed by the great bulk of our future aims. Should conflict arise between them, there is no question but that I must abandon my present aims.

This is not to say that Self-Interest Theory would require us to abandon our present aims as much as Utilitarianism would. For there are likely to be fewer conflicts between our own aims throughout our lives than between our aims and the aims of others. However, as in Williams' argument against Utilitarianism, it is not simply the sacrifice of a given project that is at issue. It is the question of what it is like to adopt a standpoint toward one's life wherein any project can be canceled or abandoned for the sake of future utility. This is to adopt a standpoint outside any particular time in one's life. It is to stand above one's life, so to speak, as if one could look down at it and arrange all the pieces that compose it. It is to make one's will subject to the utility calculus, to channel it through a calculus mechanism entirely external to one's own present projects, aims, values, and commitments.

This issue of channeling one's will through an "objective" mechanism is one that Williams raises in the context of discussing impartial moral viewpoints. He attempts to show that how an agent thinks about an action changes his relationship to the action, and changes the
meaning of the action itself. Williams offers an example described by Fried in which a man must choose which of two persons he will rescue in some tragedy, for he will only be able to rescue one.\(^{32}\) One of the persons happens to be the rescuer's wife. Fried justifies our intuition that the man ought to save his wife by arguing that it doesn't constitute undue bias for the man to choose to save his wife in that "the occurrence of the accident may itself stand as a sufficient randomizing event to meet the dictates of fairness, so he may prefer his friend, or loved one".\(^{33}\) In fact, this justification only holds "where the potential rescuer holds no office such as that of captain of a ship, public health official or the like".\(^{34}\) Thus the decision to save one's wife is channeled through abstract principles of fairness in order to justify one's action. Of this Williams writes:

...surely this is justification on behalf of the rescuer, that the person he chose to rescue was his wife? It depends on how much weight is carried by 'justification': the consideration that it was his wife is certainly, for instance, an explanation that should silence comment. But something more ambitious than this is usually intended, essentially involving the idea that a moral principle can legitimate his preference, yielding the conclusion that in situations of this kind it is at least all right (morally permissible) to save one's wife....But this construction provides the agent with one thought too many: it might have

\(^{32}\) Williams, 1981 (a), p. 17.


\(^{34}\) Williams, 1981 (a), p 17, quoting from Fried, p. 227.
been hoped by some (for instance, by his wife) that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this kind it is permissible to save one's wife.\textsuperscript{35}

It is the rerouting of one's motives for action through the impartial system with which Williams is concerned. I suggest that when our motives are channelled through a maximal satisfaction calculus involving the desires either of a number of persons, or merely of oneself, one effectively undermines the grounding for his actions in his projects and commitments. There is a real sense in which one cannot be "in" any given project if he is a Self-Interest Theorist. For rationality demands that the agent be "above" the project, and be ready to cast it aside if it does not maximize. There is a real sense in which the agent motivated by Self-Interest Theory cannot take a project seriously as a commitment at all.

I am not arguing that any method of rational assessment of aims alienates one from the experientially present and from one's projects. I believe the case to be quite otherwise. In a sense the Self-Interest Theorist is not even talking about any sort of assessment of aims to begin with. For Self-Interest Theory relies on a process that is simply based upon a summation of aims. It does not involve a deliberation either between conflicting aims or reasons, nor does it involve deliberations about the value of a given aim. On the

\textsuperscript{35} Williams, 1981 (a), p. 18.
contrary, it demands a standpoint beyond any substantive aims in order that all the particular aims can be summed in the maximizing equations as mere preferences. Were one to deliberate about his aims he would do so on the basis of what he at that moment took to be true and valid, for his most current assessments would be the best information at his disposal, or at least that which he must consider as being best. In this sense, then, the standpoint of rational deliberation is the present. But more needs to be said about this.

It is perfectly consistent for an agent to view his life from the present and yet still make criticisms of his projects and commitments. For the agent can use the beliefs and reasons that he has at his disposal at present to criticize and evaluate various of his present projects. What is more, when he assesses his aims he can only make use of those beliefs and cognitive skills which he has now. The agent may have certain beliefs which he believes will change--but again, he believes that these will change based upon the information that he now has.

The agent may decide to make provisions for aims which he anticipates he will have in the future, but does not now have. This may be so even if making provisions for the satisfaction of these future aims at the time of their occurrence means the
frustration of certain of his present aims. One might simply be operating under the aegis of a more important present project which, within a given domain, seeks to maximize satisfaction of as many of the preferences he will have over the course of his life as possible, even at the price of sacrificing some present projects. Thus, some component of the true or best theory of practical rationality may resemble Self-Interest Theory in seeking to maximize preference satisfaction within some limited domain. This is an option distinct from having Self-Interest Theory as a general theory and having all of one's aims be subject to the maximizing equations. Even if one gives great weight to certain aims, so long as they are simply fed into the maximizing equations like any other preference the sort of alienation as that which I have been discussing will arise. It is not so much a question of whether one will be able to complete his projects as a question of what his relationships to those projects is.

These points can be illustrated by the following example from Nagel:

Suppose for example that [an individual] now believes that in twenty years he will value security, status, wealth, and tranquility, whereas now he values sex, spontaneity, frequent risks, and strong emotions. A decisive response to the situation could take either of two forms. The individual may be strongly enough convinced of the worthlessness of his inevitable future values simply to refuse them any claim on his present concern. He would then regard his present values
as valid for the future also, and no prudential reason would derive from his expected future views. On the other hand he might treat both his present and future values like preferences, regarding them each as sources of reasons under a higher principle: 'live in the lifestyle of your choice'. That would demand of him a certain prudence about keeping open the paths to eventual respectability. In either case, his position would be formulable in terms of timeless reasons.\(^{36}\)

I have argued so far that the second strategy alienates us from our projects. This seems particularly so in the case of projects which we take to involve moral judgements, or, quite simply, values which we take to be true. For as Parfit argues, we hold values and moral judgements to be true or correct based upon our best assessment, based upon the best reasons that I have in my possession (now).\(^{37}\) If I can predict that I will have values in the future that conflict with those that I have at the present, then this requires some form of explanation if I take those values to be grounded in moral judgements rather than as being mere preferences.

This point is more straightforward when we think of the more general case of beliefs. As Parfit writes, we cannot say "Q is true but I do not now believe Q".\(^{38}\) So in our example, if the person making his decision were to take his future predicted aims to be based upon sound moral judgement, he must

\(^{36}\) Nagel, 1970, p. 74 ff.

\(^{37}\) Parfit, 1984, Book II #60. This can be seen clearly via the Russian Nobleman case. See Parfit, 1973.

\(^{38}\) Parfit, 1984, p. 154.
already have come to hold them as valid moral aims. He could explain the unclarity of this point in his own mind, or in his failure to fully actualize these values in his present life, in terms of his youthful naivete. On the other hand, if he takes his present values to be the most true or best justified values, he may form some theory about moral degeneration in later life.\textsuperscript{39} He may take special precautions against the possibility of his acting on such values in the future. For as Williams writes: "The question must arise, how prediction is, [in this case], related to acquiescence".\textsuperscript{40} He will either discount these future values, or act on them. But in either case he will have reasons now for believing that the morals which he so acts as to serve are based upon a proper moral judgement.

The other option is to treat all values as preferences. Here preferences are aims and desires which I do not necessarily have now. Nor need I see any reason why I ought to have them--though I may see things and events which will cause me to come to have them in the future. Thus these aims needn't be connected to me either by my experiencing them currently as a desire, nor in my presently having a reason to pursue them. However, I may have the purely general, abstract reason to maximize the satisfaction of all my desires over the

\textsuperscript{39} Williams, 1981 (a) p. 10.

\textsuperscript{40} Williams, 1981 (a), p. 9.
course of my life. This reason is peculiar in that it endorses an entirely second order project gaining content only from my first order desires. It asks me to make a judgement about what I ought to do from a standpoint outside of any particular aim regardless of what attachment I might feel to certain particular aims. It is a conception of how one ought to act, writes Williams, which "seems...to imply an external view of one's own life, as something like a given rectangle that is optimally to be filled in".41

That is why the preference maximization model may be productive of the kind of alienation we have been discussing in a way which all those theories involving rational criticism of one's aims may not. For such criticism and assessment occurs from the standpoint of one's best current understanding of the situation—and, more specifically, what moral judgement one now makes. As Williams writes: "The correct standpoint on one's life is from now".42 It is possible, then, that one may offer rational assessment of one's own aims without adopting a standpoint outside of any particular moment in the way that Self-Interest Theory asks us to.

This argument is very close to that made by Parfit in

41 Williams, 1981 (a), p. 12.

Reasons and Persons. However, it differs in this key respect: so far I have said nothing about whether reasons are temporally relative or temporally neutral. I have merely been making a point about the phenomenology of practicing, that is—living in the theoretical world of, a certain type of theory of rational action. The aim of the argument is the same as the argument against Utilitarianism; it is, to paraphrase, not "Do you agree with Self-Interest Theory's answer?", but "Do you really agree with Self-Interest Theory's way of looking at the question?". As we shall see in Chapter III, there is a strong reply which can be made to such an objection to Self-Interest Theory.  

In this argument I have attempted to demonstrate that an assumption common to Utilitarianism and Self-Interest Theory is in fact false. The assumption is "that processes of thought are transcendental to experience and do not actually take up any psychological room". I have attempted to demonstrate that for Utilitarianism as well as for Self-Interest Theory there are real and important phenomenological consequences of adopting either the impersonal standpoint or

---


44 Remarks like those made by Williams, 1981 (a) at Section II pp. 10-13 are suggestive of the argument that I have made here at length, but I do not believe that Williams fully articulates such an argument.

the temporally neutral standpoint; that, in Williams' words, "to think in one way rather than another about what to do is to be empirically different, to be a different kind of person...". The aim was to "show up in greater depth what would be involved in living with these ideas". I have done this through considering the fundamental role that projects and commitments play in our lives, and the fundamental impact which the perspective that we take on them can have in our lives. Williams puts the point as follows:

Life has to have substance if anything is to have sense, including adherence to the impartial system; but if it has substance, then it cannot grant supreme importance to the impartial system, and that system's hold on it will be, at the limit, insecure.

I have argued that this point applies to Self-Interest Theory as well as to Utilitarianism. We cannot regard our values as merely summable preferences and at the same time take them seriously as values. I believe that this argument stands on its own. However, in Chapter 2 I further will develop the point of this argument in terms of considerations surrounding the agent's abilities and motives to maximize. That argument will support the argument of this chapter by offering a closer analysis of the motivation for action in certain problematic

46 Williams, 1981 (c), p. 51-52.
47 Williams, 1973, p. 78.
cases, and then attempting to generalize that result. In the course of the argument of Chapter 2 a further sort of alienation caused by Self-Interest Theory as a standpoint of rationality will be discussed. This alienation is even more profound than the alienation discussed in this chapter.
CHAPTER 2
Introduction

The Self-Interest Theory Maintains that the agent ought to maximize the satisfaction of preferences over the course of his entire life, treating all times of his life equally. However, actual human agents suffer from weakness of the will or "time preferences". That is, they are often led astray from their larger goals by the temptation of some immediate payoff. In order to maximize they must overcome weakness of the will. In what follows, I will examine a method for overcoming weakness of the will discussed at length by Elster and others: precommitment.¹ I shall attempt to demonstrate the following. When agents suffer from weakness of the will they do not maximize, and when they overcome weakness of the will through precommitment they do not do so in order to maximize. They do not seek to maximize preference satisfaction through precommitment because to do so would involve a kind of alienation more profound than that discussed in Chapter 1. I will then attempt to apply this result more generally arguing that if agents are not motivated to maximize preference satisfaction when they precommit, which is ostensibly the paradigm case of successful maximizing rationality, we have little reason to believe that they are

¹ Elster, Jon, 1984, Chapter II.
motivated to do so elsewhere.
I. Time Preferences, Weakness of the Will and Precommitment

It is fundamental to self-interested rationality as it is generally understood among philosophers and economists, and to Elster's theory of rationality in particular, that time preferences are irrational. In *Ulysses and the Sirens*, Elster assumes this view in order to make his arguments.\(^2\) By time preferences he means the favoring of one period of time over another in our allocation of resources merely because it is that particular period of time rather than because it is believed that certain events or conditions will obtain during that time. Typically, this would involve allotting a disproportionate share of resources to the near future. By "disproportionate" I mean a proportion over and above what is warranted on the basis of discounting preferences for future possibilities due to the uncertainty of future events.

An example of time preferences is that I might, for instance, have a large piece of cake now, despite the fact that the pleasure that I gain from it will be outweighed (from a temporally neutral point of view) by my future suffering from the indigestion which it is bound to cause. Nonetheless, I eat the cake because I favor the present and nearer future

\(^2\) Elster, 1984, pp. 66-67 and footnote 61 on the same pages.
over the further future with respect to my utility consumption. In such a case one frequently feels regret for one's past choices. One natural way to describe this situation is to say that one fails to resist the temptation of an immediate payoff, and thus the failure is called "weakness of will". In fact, this is Elster's interpretation of time preferences. He writes: "This interpretation of time preferences as weakness of the will is crucial to my argument."³ Elster concludes that:

Qua rational beings, and abstracting from the problem that we know that but not when we shall die, we want to allocate our resources equally over time: A year is a year is a year: there are no grounds for preferring the present over the future simply because it is the present.⁴

Time preferences, then, are construed as a flaw in the rationality of the agents. A large part of Elster's task in Ulysses and the Sirens is to discuss techniques by means of which agents of limited rationality can overcome this flaw.

Means of overcoming time preferences must be such as would allow the agent to overcome the temptation of some earlier smaller reward for the sake of gaining some larger, later reward. He must be able to wait in the face of temptation. In other cases, one may prefer A to B, and prefer

³ Elster, 1984, p. 67.
C to A. If one can proceed from A to C only through B, the rational agent ought to choose B over A in order to get to C. One moves one step back in order to move two steps forward, so to speak. Elster calls this ability, which allows the agent to endure temporary setbacks for the sake of greater future rewards "global maximization".  

"Local maximizers", on the other hand, will always maximize utilities in the present moment, despite the consequences for future utility consumption. Elster takes the generalized capacity to be global maximizers to be the hallmark of rationality.  

Weakness of the will, or time preferences, are then a failure to maximize globally. In what follows I will discuss precommitment as an indirect strategy which allow agents to maximize globally despite the pull toward being mere local maximizers.

Given what we have said about time preferences, a rational agent may come to recognize his own temporal biases and try to overcome them. If one wishes not to succumb to his own preferences based upon time discounting, the situation is one of conflicting preferences. I am here concerned with

---

5 Elster, 1984, p. 9.

preferences that change across time within a single individual. The relevant cases are not simply those in which a person's tastes or opinions change in a straightforward way. Rather, they are of the sort such that at some time, t1, an individual has some preference which he wishes to satisfy at some future time, t3. Nonetheless, he, realizes at t1 that at some intervening future point, t2, he may succumb to the temptation to satisfy a preference which he has at t2. If he succumbs to the temptation at t2 he will no longer be able to satisfy the original preference that he had at t1 (and will still have at t3). The agent may fail to resist the temptation at t2 even though he knows that he would be better off overall (maximize globally) if he were to resist that temptation.

An example of this sort of difficulty is that an agent, call him John, might now want to work at the office late tonight. At the same time, however, John might realize that at five o' clock, when his coworkers leave to join family and friends, he too will prefer to leave the office. Here, it may be the later preference that comes to be regarded as somewhat troublesome, for it may cause John to fail to get his work done in a timely fashion. Therefore John's overall interests--his interests across time--would be best served by overcoming the later preference. However, the phenomenon could be structured the other way around. For instance,
George might realize now that if he frivolously spends his entire paycheck he will later come to regret it, preferring that he hadn't done so. He would be better off overall if he were thrifty with his money. Nonetheless, it may be George's strongest preference now to blow all his money. Here it is the earlier preference that is problematic.

Given that an individual can recognize this as a possible configuration of preferences in his life, he might take steps to prevent himself from acting on his troublesome preferences in order to maximize his overall utility. In the two examples given above, one might engage in what Elster calls "precommitting" oneself. In the first case this could be done by John pulling down his office shades so that he didn't see his coworkers leaving, and thus wouldn't feel the temptation to leave himself. In the other case, George might ask to be paid on a daily basis so that he cannot spend two weeks' wages on an impulse. In these ways, the agents keep themselves from acting on their troublesome preferences so that they can do what they most prefer to do overall.

Taking a closer look at precommitment, we can see that it involves setting up a causal process outside of ourselves. Elster writes:

7 Though I am more interested in the phenomenon itself, rather than in Elster's characterization of it, I will focus on his account for the sake of convenience.
Our intuitive notion of what it is to bind oneself seems to require that we temporarily deposit our will in some external structure; that we set up some causal process in the external world that at some time returns to its source to modify our behavior....We are talking about controlling the controller in order to control oneself.\(^8\)

Elster proceeds to "sketch in rough outline the contours of the notion" of precommitment in terms of five criteria.\(^9\) However, he does "not here pretend to give necessary and sufficient conditions" for precommitment in terms of these criteria.\(^10\) They serve only as "first approximation" of such criteria as would be used to distinguish precommitment from various related phenomena.\(^11\) In order to better understand Elster's notion of precommitment, I will list and briefly comment upon these criteria.

The criteria are relatively straightforward:

(i) To bind oneself is to carry out a certain decision at time \(t_1\) in order to increase the probability that one will carry out another decision at time \(t_2\).

(ii) If the act at the earlier time has the effect of inducing some change in the set of options that will be available at the later time, then this does not count as binding oneself if the new feasible set includes the old one.

\(^8\) Elster, Jon, 1984, p. 43.

\(^9\) Elster, 1984, p. 46.

\(^10\) Elster, 1984, p. 46.

\(^11\) Elster, p. 39.
(iii) The effect of carrying out the decision at t1 must set up some causal process in the external world.

(iv) The resistance against carrying out the decision at t1 must be smaller than the resistance which would have opposed the carrying out of the decision at t2 had the decision at t1 not intervened.

(v) The act of binding oneself must be an act of commission.\textsuperscript{12}

These criteria are fairly straightforward, but a few points should be noted. With respect to (ii), the point is that "some options must be excluded if we are to speak of precommitment".\textsuperscript{13} For, otherwise, any case of investing resources, even when they could be withdrawn at any time with no loss to the feasible set (and possibly some profit), would count as precommitment. This would not seem to represent a truly binding act. Criterion (iv) simply makes reference to the fact that if acts of precommitment were more difficult to initiate than directly acting so as to achieve one's goal, there would be little or no interest or value in such acts. There would be no rational motivation to engage in precommitting under such circumstances. With respect to criterion (v), Elster notes that the distinction between acts of commission and acts of omission is notoriously difficult to make. Nonetheless, he wants to stress that precommitment is

\textsuperscript{12} Elster, 1984, pp. 39-47.

\textsuperscript{13} Elster, 1984, p. 42.
a deliberate rational strategy. An agent purposely binds himself so that he must forgo certain actual preferences in order that he might enhance his ability to satisfy other preferences.

Elster provides the following examples of precommitment:

In order to stop smoking it is standard practice to set up some causal machinery that will add force to your inner resolution: [1] to tell your friends about your intention so as to invite their sarcastic comments if you are backsliding; [2] to go for a walk in the mountains so as to make cigarettes physically unavailable; [3] to cross the street when you see a tobacco shop further on so as not to be exposed to the sight of cigarettes; [4] to take cold showers in order to strengthen your will power; [5] to undergo hypnosis in order to induce aversion to tobacco; [6] to make yourself believe that cigarettes mean certain death within five years.\(^1\)

This passage invites comment because it contains examples which involve several different methods of precommitment. [2] involves a means of physically preventing oneself from acting on one's own preferences in a fairly straightforward way. [1] seems to involve binding preferences to smoke with other preferences, namely preferences not to be ridiculed, so that they cannot both be satisfied simultaneously. The preference to smoke and not be ridiculed is no longer an option. [6] does the same with respect to our preference not to die, or cause ourselves to die. Here, however, the mechanism is activated differently, by causing ourselves to

\(^{14}\) Elster, 1984, p. 37.
have a belief, rather than causing others to have a belief. [4] seems to involve acting so as to stimulate desires that are favorable to overall utility maximization.

Only [3] and [5] involve acting on the offending preference itself. [3] involves taking conscious steps to avoid the stimulation of the problematic preferences. It is for this reason that the ex-smoker might avoid tobacco shops. [5] involves a straightforward assault on the desire itself. Though these two techniques can be regarded as distinct items, in particular cases they may blend into a single phenomenon. For instance, if the ex-smoker avoids stimuli which cause him to desire to smoke for long enough ([3]), the tendency to have that desire may itself diminish, or disappear entirely (similar to [5]). To a lesser extent, this holds for the other four strategies as well. Maintained for a lengthy enough period of time, any one of the first four techniques discussed ([1], [2], [4], [6]) might achieve the indirect result of encouraging or allowing the offending desire to dissipate since that desire would not be reinforced through the agent's acting on it. This would be particularly likely to occur in the case of such techniques as [1] where the desire itself may bring up unpleasant associations. These results may be either an intentional or an unintentional by-product of these strategies.
We might consider how examples (3) and (5) fit the five criteria of precommitment in order to get a better feel for how precommitment operates. Suppose that I know a tobacco store is down the street. I know that if I walk directly by the store I will see and smell the fresh tobacco. I will be unable to resist buying and smoking some cigarettes. However, overall I believe that it would be better for me not to smoke cigarettes. It is only in such moments of weakness and temptation that I would do so. I therefore decide to cross to the other side of the street so that I will not, in a few moments, be tempted to smoke (criterion (i)). This decision requires, for its fulfillment, that I change my physical location and path of travel (criterion (iii)). I do this act deliberately (criterion (iv)). It is easier to cross the street now than it would be to avoid temptation later if I walked in front of the store (criterion (iv)). Though I have gained the option of crossing the street again and going back to the store, I have lost the option of simply turning into the store as I walk in front of it (criterion (ii)).

In the hypnosis case, I decide to undergo hypnosis now in order to decrease the probability that I will decide to smoke later (criterion (i)). If successful, this act would exclude smoking, or at least pleasure in smoking, from my future feasible set (criterion (ii)). My decision required that the hypnotist hypnotize me (criterion (iii)). It is easier to be
hypnotized than it is to give up smoking without such assistance (criterion (iv)). This decision requires that I take a positive and purposive act (criterion (v)). The remaining examples fit the criterion of precommitment in a fairly straightforward and analogous way.

It seems that in many cases the elimination of the problematic preference would be the ideal resolution of the problem. For otherwise, one might risk the constant battle to overcome temptation, a battle which is likely to be lost at least some of the time. Stoic and Buddhist philosophies have long advocated the reduction of such problematic desires as would frustrate well-being, and there are many more mundane bodies of knowledge which advocate such "technologies of self-management".\(^{15}\) Thus, methods of precommitment may also be, either intentionally or unintentionally, methods of character planning.\(^{16}\)

\(^{15}\) This term is taken from Elster, 1983, II.3.

\(^{16}\) Elster writes: "The goal of precommitment may either be to make certain options permanently unavailable, or to make them permanently undesirable as a result of their having been temporarily unavailable. The latter would be precommitment in the service of character planning..." (Elster, 1983, p. 114, footnote). He does not mention the obvious possibility that precommitment may result in changes of character that are an incidental byproduct of the strategy rather than its goal.
II. Precommitment and the Motive to Maximize

Precommitment is, then, a tool for overcoming weakness of the will, and it is precisely this weakness that a human agent will have to overcome in order to approximate the ideal of maximization of preference satisfaction. In what follows, I will examine a problem with precommitment, or, more exactly, a limitation on it. I will not seek to cast doubt either upon the existence or the usefulness of the behavioral technique known as "precommitment" in overcoming weakness of the will. I will, however, criticize the way in which it has been characterized. Specifically, I will examine the motives for precommitment, and argue that the maximization of utility is not the reason why a given agent precommits, except, perhaps, in a small number of cases. That is, maximization of preference satisfaction serves as a reason to precommit only as one sort of reason amongst a large variety of particular reasons for which an agent might precommit. In short, maximization is not the privileged motive for precommitment. If this argument succeeds, then it follows that if an agent overcomes weakness of the will, his motive for doing so will typically not be maximization. Therefore, it seems that, at least with respect to this behavioral technique, either agents do not overcome weakness of the will and therefore do not maximize; or they overcome weakness of the will, but only for
some other particular reason which, typically, will not be utility maximization.

I will attempt to make this point by offering an example of precommitment and then considering how best to interpret what the agent's motive to precommit could be. The sort of case that will best lend itself to my argument involves a series of acts of precommitment. Each in this series of acts of precommitment will, either directly or indirectly, alter the character of the agent. I will argue that, ordinarily, an agent's motive for engaging in each of these acts of precommitment will not be utility maximization. For he would not accept the series even when it is true that, had he undergone these series of precommitments under certain conditions of ignorance, it would have maximized utility. He will not accept the series as a whole as to do so merely for the sake of utility maximization would involve a profound sort of alienation. I then go on to argue that, if utility maximization does not motivate him over the series of acts of precommitment taken as a whole, then it is not sufficient to motivate the agent in each of the particular instances of precommitment. I will then argue that these results apply generally to the motives for individual acts of precommitment whether or not they involve character change. To make this argument clear it is best to start with an example.
Scenario 1

Consider the case of some individual agent, Bill. Bill is an investigative reporter and, given the sort of story that Bill typically investigates, he is generally to be found fraternizing with some of the more disreputable characters in town. It is a great boon to Bill's endeavors to gather information that he has many of the same habits as those from whom he frequently seeks to gather information. In short, Bill tends to drink and smoke excessively, and has rather loose sexual morals. However, one day, while driving home from one of his journalistic forays Bill collides with a large tree. Though neither he nor any one else is seriously injured, Bill is greatly concerned. The accident raises certain problems with his boss. If Bill's boss forms the impression that Bill is a drunk Bill could lose his job. Bill loves his job. He cannot let this happen.

Bill tries to return to his work as before, with one exception: he now wishes to drink only in moderation, rather than somewhat recklessly as he formerly did. Unfortunately for Bill, he is a man of weak resolve in this matter. Once Bill enters a smokey bar room or meeting room, and hears the clink of ice hitting the bottom of a glass and smells the aroma of whiskey as it splashes over the ice, it is all over for him. He's on his second double before he knows what happened. Bill decides that he must not drink at all, for he
cannot drink in moderation. He, therefore, takes a drug which will make him violently ill if he drinks. Bill has precommitted himself against drinking.

But going cold turkey is tough for Bill. He starts reading a lot of self-help books about alcohol abuse and they recommend that the alcoholic seek assistance through spiritual or religious practices. Bill strongly dislikes religion, but he believes that if being a practicing churchgoer would ease the problem of staying sober then it would be worth it. However, Bill knows that, come Sunday, he will undoubtedly oversleep, as he always does, unless he takes some precaution against it. Bill joins the altar society. By doing this Bill ensures that he will face great social embarrassment if he does not attend church regularly. Bill has bound himself to attend church.

While he is involved in with the altar society Bill meets Betty. In a short time they become strongly attracted to one another. Betty does not approve of Bill's aforementioned sexual license. Bill, concerned to curry favor with Betty and build a relationship, decides to forswear certain sexual freedoms which he previously enjoyed. However, as always, decision is one thing and action is another. In order to avoid infidelity Bill no longer frequents the bars and clubs where he would be likely to become involved in sexual
liaisons. He avoids them altogether in order to avoid succumbing to the much stronger temptation he would face were he to encounter one of his old lovers under such circumstances. Again, Bill has precommitted to overcome weakness of the will.

We can imagine Bill and Betty moving on toward marriage. Perhaps Bill finds it hard to do the sort of investigative work he used to do in view of the changes in his behavior already discussed. It is simply harder to fit in with the scene. He seems to be enjoying it less. In fact, he decides to become a sports writer. It is much more satisfying work given his current social position, and is a job of which Betty more readily approves. We can imagine Betty and Bill getting married, moving to the suburbs, settling down to raise kids, and so forth.

In this story an individual, Bill, has greatly changed his character. The overall change in his character was, in large part, due to a series of small changes in his character which resulted from acts of precommitment. The story involved rather dramatic incremental changes as it was told here. However, this obvious kind of change is not necessary for the argument which I wish to make. It is only because I had to tell the story wherein significant character change occurred while at the same time making the story short enough to be
appropriate to this thesis that the changes had to be portrayed as abrupt. Many other sorts of stories could have been told which exhibited the same phenomenon in ways more complex, subtle, and interesting than the one described here. It was the constraints of the narrative form, rather than the phenomenon, which makes the character change seem somewhat contrived.

Having laid out the story, I would now like to assume that the choices that Bill made were the best ones that he could have made given his circumstances. That is, I want to suppose that Bill could have done no better in terms of seeking to maximize his own satisfaction. To the extent that the concept makes sense (and I am not admitting that it is fully coherent) Bill has maximized total utility both in terms of each decision to precommit which was made and its immediate consequences, and for his life as a whole. In choosing the path of reform, Bill's life has gone as well as possible.

Scenario 2

Given all of these stipulations, I believe that the following can be true in cases such as this one. Suppose that the world is exactly as the world described in Scenario 1 in all respects save one. A short time after Bill has wrecks his car, while he is contemplating what he should do and whether
he should take the pill which will cause him to be ill if he consumes alcohol, a visitor comes to Bill's house. The visitor is a sociologist who happens to have a super computer (or perhaps he is a Cartesian evil demon pretending to be a sociologist with a supercomputer). The visitor tells Bill that he has fed the computer all the relevant data concerning Bill's life. He tells Bill the following. "Had I not arrived to make this announcement, and if you took that pill there is a high degree of probability that your life would unfold in the following way". The sociologist then tells Bill of the series of events in Scenario 1. Without the visit from the sociologist, Bill's life would have taken a course rather different from the one it has heretofore. Now conditions have changed from the original case in that Bill has this new information about what he would have done if he hadn't heard the sociologist's prediction. Would Bill make the same decision as in the first scenario?

I think that it is quite intelligible to say that Bill would not make the same decision as he did in Scenario 1. In fact, in the sort of case described here it seems likely that Bill would act differently if given the new information. It is entirely likely that the hard-boiled investigative reporter would find the prospect of living as a mild-mannered suburban sports reporter with his wife, two kids and a dog to be a dreadful prospect, and a fate to be avoided if at all
possible.

The only difference between Scenario 1 and Scenario 2 in the initial stage is the information provided to Bill. Bill decides differently in the face of this information. But this is not an ordinary case where further information yields a better decision. For we might imagine that Bill in scenario 2 fares less well in terms of happiness or preference satisfaction than Bill in scenario 1. This seems entirely intelligible. The life of the investigative reporter in 2 is unlikely to promote more happiness than the life of the reformed man in scenario 1 in that essentially that same life lost out in Scenario 1 over the option of the reformed life. Perhaps the unreformed Bill in Scenario 2 will take some satisfaction in not "selling out" in Scenario 2 as he would have in Scenario 1, but the decision to avoid being what he never wanted or intended to be is likely to promote a small bit of satisfaction at best. The reformed life in Scenario 2 may seem the worst option of all in that Bill foresees its outcome and is jaded with regard to the acts of precommitment which would be open to him. With each decision he might feel that he is "selling out". The attempt to reform, if it succeeded at all, would probably produce in Bill feelings of contempt, and perhaps loathing, for himself and his life. The interesting question is why these results seem right, and in explaining that we will see more clearly why Bill would be
motivated to refuse to reform in Scenario 2.

In Scenario 2 Bill is given information about what his life might have been. He looks at the life that would have been the culmination of his efforts at personal reform and is filled with disgust. Bill takes such a life to be boring, petty, and horribly ordinary. Williams writes that regarding such changes in one's orientation toward life:

> We are here concerned with distinctive and structured patterns of desire and project, and there are possible psychological changes in these which could be predicted for a person and which would put his future after such changes beyond his present interest. Such a future would be so to speak, over the horizon of his interest...\(^\text{17}\)

I have constructed a case in which a person is not only disinterested in these potential character changes, but is, in fact, averse to them when he sees them in their totality.

I have not made Bill's objections to the reformed Bill of Scenario 1 specifically moral objections. In fact, if morals come into play at any point it is in Scenario 1 in which the decisions of Bill, later in his life, are more likely to have a moral motivation. This argument is not, therefore, a rehearsing of the argument about the nature of judgement and values across time as given in Chapter 1. One may construe

\(^\text{17}\) Williams, 1981 (a), p. 9.
Bill's choice in Scenario 2 to be based upon values, but this would involve a broader notion of values than is usually encompassed by the notion of moral values. Even if the case of Bill is interpreted as essentially similar to the argument about value judgements given in Chapter 1, then this case will broaden the range of application of that argument. This would merely be another way of reaching the same conclusion: contrary to Self-Interest Theory, the standpoint of rationality is from here.

In this case the problem is that Bill simply does not identify with the reformed character of the agent at the end of the story in Scenario 1. He does not identify with the interests of that character as being his interests. In fact, his interests (present interests, interests that he possesses) are in conflict with the interests of this other character. The bare knowledge that it would be him is not enough to make the path of reform seem rewarding. In fact, were Bill to attempt to head down that path it would be in the knowledge that he is working toward becoming something that he doesn't want to be. If this knowledge didn't make it impossible for Bill to pursue this path it would certainly alter his perceptions and experiences such as to make the path manifestly unpleasant, and would tend to alienate him from his emotions and convictions as well as his acts. Bill would be acting directly so as to become someone whom he did not want
to be. Surely this would promote a deep alienation from one's very character. Given this, it seems likely that the path of reform in Scenario 2 would promote not only more unhappiness for Bill than the path of reform in Scenario 1, but also more unhappiness than the path of non-reform in Scenario 2. Thus, a difference in perception causes a difference in the amount of satisfaction that identical states of affairs would promote on otherwise identical agents.

After explaining that a certain future of oneself could be beyond one's "horizon of interest", Williams writes that, on the other hand, "of course if the future picture could be filled in as a series of changes leading from here to there, [one] might recapture an interest in the outcome".\(^{18}\) I have presented possible future changes in Bill's character as resulting from a series of changes. I have argued that he still would be averse to it. I claim that some changes in one's character can be of a sort such that one's future desires and preferences are unintelligible or unacceptable from the present, except as they are designated abstractly as preferences. If future aims are insufficiently similar to present aims there is no reason to believe that we could comprehend or appreciate them, let alone approve of them. Bill may appreciate them abstractly and externally as "aims which could give me pleasure", just as one may understand that

\(^{18}\) Williams, 1981 (a), pp. 8-9.
donating to Oxfam may contribute to the benefit of some unknown person. But this is not the same as claiming them as one's own in a deep sense.

In order for Bill to claim these interests as his own, internally, so to speak, rather than merely stipulating them as potentially his (in order not to be alienated from them), he must be able to tell some story about how his present self connects up with that future self so that these aims are intelligible to him now. We might call this a "narrative of intelligibility". However, in the story I have told, ex hypothesi, there is no such narrative. Or, rather, there is only one such narrative open to him, and that is one of the most minimal sort. For having been given full information about what would lead him to reform his life, Bill still finds it an odious outcome. His judgement about what sort of life is best is simply not transformed by this information.

Were Bill to choose the path of reform in Scenario 2, he could regard these preferences only as his future preferences, and the path of reform as possibly a way of maximizing his preferences. "Bill's preferences" would here be an entirely abstract designation adequate only to this narrative of intelligibility: that Bill ought to reform because it would maximize utility for him. I submit that this narrative would be entirely incapable of producing for Bill in Scenario 2 the
motivation that Bill had in Scenario 1. The path of reform would remain quite unsatisfying, because the only narrative of intelligibility open to him would be insufficient. In Scenario 2 Bill would be worse off overall, in terms of any intuitive notion of well-being, if he attempted reform along the lines he would have taken in Scenario 1. There is good reason to conclude that it would be rational for Bill to attempt to abstain from the attempt to reform in Scenario 2.

If one were motivated by the drive to maximize preferences in this case one would have an extremely peculiar relationship toward one's character. For from an ideal standpoint arbitrarily favoring some preferences over others could not happen. This includes favoring present preferences over future preferences. Therefore, one could have no concern as to the shape or nature of one's character except as to whether it maximized preferences. Hence, ideal rationality prescribes utter unconcern for the content of one's character. One ought to be no more attached to his character than he is to a sum in a bank account. Character would be, in effect, readily convertible capital to be utilized in whatever way would maximize under current market conditions. At this point, the I which oversees these transactions recedes to a place outside of any action occurring in the lifeworld. It claims this life as its own not because it is embodied in that life and the substantive aims and commitments which
distinguish it from all other lives. Rather, it claims this life because it is the one over which it has control and within which it can make various pleasures and pains occur. It is not attached to that life by any valuing activity, such as bearing love toward it, exhibiting fidelity to it, or by identifying with it. Rather, it bears a managerial relationship toward that life. It is as if one would say "This body, this personality, this character are the means by which I derive my pleasure". The disregard for the content of first order values and aims is striking.

The discussion of Bill's motivation in Scenario 2 also tells us something about Scenario 1. That is, if the motive of maximization was not sufficient to properly motivate Bill in Scenario 2 then it could not have been sufficient for Bill in Scenario 1. That is, if Bill's reason for precommitment at every stage in Scenario I had been maximization of preference satisfaction, and had this produced the best outcome, then this should have been a sufficient reason to precommit in Scenario 2 and would have produced the best outcome when the same series of decisions was considered as a whole. But maximization would not be sufficient or proper motivation to precommit in Scenario 2 and would not produce the best outcome; therefore, utility maximization could not have provided Bill with sufficient motivation in Scenario 1 nor could it have produced the best outcome.
If we suppose, on the other hand, that Bill in Scenario 1 was acting not for the sake of maximization of preference satisfaction, but for certain particular aims and values, and from the perspective of an agent with those values, then the pieces of the puzzle fall into place. For each step in the progression of precommitments Bill was doing what he thought was best at that particular time in terms of his particular aims and purposes. That is why the move from one stage to the next is intelligible to him. His aims for the future were, at each point, explicable in terms of aims and projects that he had in the present. There was a rich interconnection between the present and the future. In Scenario 2 Bill was given information which forced upon him an external perspective on his potential future acts wherein he could appreciate his future aims merely as abstract preferences. As we saw, this conception of his own good and self-interest was insufficient to motivate Bill to reform through precommitment.

If maximization is not the motivation to precommit in cases such as this one, where there is a marked character change, there is reason to doubt that it is the motivation in other cases of precommitment involving lesser character change. For from the agent's standpoint he cannot tell when one act of precommitment is part of a larger series that will radically change his character. Therefore, his reasons for action are likely to be insensitive to such differences. In
the case of Bill, I argued that in Scenario 1 his motive was not to maximize, and in this case the agent was unaware that each decision to precommit could be seen as a part of a larger series that would radically change his character. Therefore, it would seem that the typical motive to precommit is not mere maximization even when the agent does not foresee the particular instance of precommitment as being a part of a larger series that will dramatically affect his character.

This analysis does not apply only to cases of precommitment involving character change. The character changes which resulted from precommitment in the preceding example were largely unforeseen side effects, rather than intended consequences. As the story was told, Bill set out to change his behavior, not his desires or character. His motivation was largely insensitive to the changes in his character that his acts might produce. Therefore, the motive to precommit would not be the maximization of preference satisfaction regardless of whether the act of precommitment was intended to produce character change or not. The lesson derived can be applied to cases of precommitment generally. Thus, from the standpoint of the agent's reasons for action, when he precommits, when he chooses to forgo some future reward for the sake of some present benefit by binding himself, he does so for particular reasons. This is not to say that he may not, on occasion, be motivated to maximize
preference satisfaction, but it would be just one reason among many for the agent to precommit.

We have thus obtained a curious result. Precommitment was introduced as a means by which the agent could overcome weakness of the will and maximize utility. Yet if my interpretation of the foregoing case is right, then efficacious reasons which lead actual agents to precommit will be particular aims, and only incidentally will the aim be to maximize utility. What looked like a paradigm case of self-interested maximizing rationality turned out to be something else altogether. This result supports the more general analysis of motivation given in Chapter 1. Together these arguments show that there is good reason to believe that agents neither are, nor ought to be, fundamentally motivated by maximization of preference satisfaction.
CHAPTER 3

Introduction

In Chapter 1 an argument was made against Self-Interest Theory which paralleled Williams' argument against Utilitarianism. In Chapter 2 a second argument was offered against the Self-Interest Theory which reinforced and extended the conclusions of the first chapter. In this chapter an argument will be made against Self-Interest Theory which, in part, relies on the results of the first two arguments. This third argument will also be modelled on an argument that Williams' makes against Utilitarianism. As one might expect, Williams original argument against Utilitarianism, upon which my argument will be modelled, relied in part upon the Williams' argument which was discussed in Chapter 1. In brief, the main argument of this chapter will be that Self-Interest Theory is self-defeating because it produces the kinds of alienation and other difficulties for the agent that were discussed in the first two chapters of this thesis.

Having made the argument that Self-Interest Theory is self-defeating, I will look at two objections to this argument which are given by Parfit. One objection is that Self-Interest Theory is not directly self-defeating but is merely
"indirectly self-defeating". The second objection to the argument against Self-Interest Theory is that it is merely "self-effacing" and not directly self-defeating. This argument also serves as an objection to the arguments of the first two chapters. The basis of these objections is that for a theory to be indirectly self-defeating or self-effacing is not for it to be damagingly self-defeating.

In response to these objections I will argue, based upon the results of Chapter 2, that Self-Interest Theory is more than indirectly self-defeating, and is, at least, self-effacing. I will then consider four arguments which attempt to show that if Self-Interest Theory is self-effacing it cannot be the true or best theory of practical rationality.

In the first of these four arguments I will argue that if Self-Interest Theory is self-effacing it cannot provide reasons for action, and so is not a theory of practical rationality. I will also argue that the intuitions lying behind the notion that it is rational for an agent to seek to maximize his own utility might be better interpreted as supporting a theory of the good for agents, rather than a theory of rationality for agents. Secondly, I will argue against Self-Interest Theory on grounds that challenge the legitimacy of the notion of a self-effacing theory. Thirdly, I will argue on empirical grounds that even if Self-Interest
Theory is self-effacing, it is still self-defeating. Lastly, I will argue that, given certain widely shared conceptions of moral epistemology, we could have no reason to believe Self-Interest Theory if it were self-effacing. I do not take these arguments to be conclusive, but they make Self-Interest Theory appear considerably less plausible than it originally did.
I. Whether Self-Interest Theory Is Self-Defeating

In what follows I will examine the sort of problems that Williams argues that agents will have should they seek to be Utilitarians. Specifically, I will focus on those difficulties which lead Williams to believe that Utilitarianism is self-defeating in that if agents chose to be Utilitarians, the aims of Utilitarianism would be worse achieved than if they had some other set of projects.

Much of the force that the claim that Utilitarianism is self-defeating has can be gotten from arguments of the sort given in Chapter 1 of this thesis. That is, if agents are motivated by Utilitarian reasons, such that the nature of their participation in first-order projects is governed by these reasons, then they will, so to speak, stand at one remove to these projects.

The problems arising from standing outside of one's projects was thoroughly discussed in Chapter 1. If we adopt the standpoint of Utilitarian assessment we are forced to stand outside of our projects and commitments such that we must regard our deepest moral intuitions as mere feelings. These intuitions must be regarded as being just like any other feelings of any other agent, and are not to be granted a privileged status in our motivation for action. But to live
in terms of this perspective, I argued, was to be alienated from one's moral emotions. I also argued that to be willing, at any moment, to step aside from one's deepest projects, or to take on a project that one finds repugnant, merely because the Utilitarian sums come out that way, is to be alienated from one's actions, and the source of one's actions in one's commitments. A world in which agents generally adopted this stance seems a dreary, if not nightmarish, world indeed. Given the clash between first-order projects which give our lives meaning and purpose and the disposition to maximize, there would be a greater happiness overall if agents adopted some other theory of morality. Utilitarianism seems to fail in its own terms.

In an effort to make our projects clash less with the demands of Utilitarianism, we might try to make the content of our first-order projects more congenial to the higher-order project of the utility calculus. However, the disposition that urges one to maximize utility leaves little motivational space for other kinds of dispositions. Suppose, for instance, that spontaneity is regarded as a valuable aspect of character. Williams quips: "...if spontaneity has utilitarian value, then doubtless we can organize some spontaneity".¹ That such an idea as "organizing some spontaneity" seems ridiculous drives home the point that

¹ Williams, 1973.
certain dispositions and desires cannot be effectively managed by the principles of Utilitarian maximization. If one is a
thoroughgoing Utilitarian maximizer, there is little space for spontaneity. Williams writes:

The difficulty is that such dispositions are patterns of motivation, feeling and action, and one cannot have both the world containing these dispositions, and its actions regularly fulfilling the requirements of Utilitarianism...If you want the world to contain generous, affectionate, forceful, resolute, creative and actually happy people, you do not wish it to contain people who uniformly think in such a way that their actions will satisfy the requirements of Utilitarianism.²

Thus the calculating mentality of Utilitarianism creates a certain sort of agent, and a world composed of such agents is not likely to be a particularly happy one.

The lesson here is that Utilitarianism makes the mistake of acting as if:

...the processes of practical thought are transcendentally to experience and do not actually take up any psychological room. But in fact to think one way rather than another about what to do is to be empirically different, to be a certain kind of person, and it is not possible to combine all kinds of reflection with all kinds of disposition.³

The disposition to maximize utility appears, then, to be self-defeating as it produces a kind of reflection and behavior which impedes the development of the very dispositions which

² Williams, 1981 (c).
³ Williams, 1981 (c), pp. 51-52.
produce happiness. Thus, Williams urges us that: "Utilitarianism's fate is to usher itself from the scene".\textsuperscript{4}

The parallel arguments against Self-Interest Theory are obvious. Self-Interest Theory alienates oneself from one's actions and moral emotions in the ways described in Section II of Chapter 1. If one were disposed to be a thoroughgoing maximizer, he could not, in an important sense, take his deepest projects and commitments seriously, and could not fully participate in them. This would make Self-Interest Theory self-defeating in that the alienation consequent upon thinking and behaving as a self-interested maximizer would not maximize utility as well as other sorts of dispositions would.

One might try to cultivate dispositions which did not clash with Self-interest Theory. However, this would be at the cost of having a robust character. For any rich set of desires and attitudes is liable to produce conflict over time in the ways discussed in Chapter 1, Section II. One is liable to find Self-Interest Theory a livable doctrine at the expense of losing the opportunity to have the sorts of character traits just discussed in the argument against Utilitarianism. Therefore, we have good reason to believe that Self-Interest Theory is self-defeating.

\textsuperscript{4} Williams, 1973, p. 134.
Williams concludes his argument that Utilitarianism is self-defeating with the line: "...I leave it open for discussion whether this shows that utilitarianism is unacceptable, or merely that no one ought to accept it". In other words, if utility is maximized only indirectly through actions and attitudes not at all grounded in Utilitarian conceptions of value, it is clear, even from a Utilitarian standpoint, that we ought not to accept Utilitarianism. What is unclear is whether Utilitarianism survives as a distinctive moral theory.

Williams is inclined to think that Utilitarianism does not survive as a distinctive moral theory. He writes that one may wish to have Utilitarianism retire to a more indirect level, such that one may cultivate dispositions which may not be Utilitarian-minded (such as spontaneity, generosity, forcefulness, resoluteness, creativity), but which nonetheless maximizes overall utility better than Utilitarian-minded dispositions. Of this Williams writes that there is "a comprehensible desire to leave the way open for utilitarianism to retire to a more indirect level". However:

...once that has started, there seems nothing to stop, and a lot to encourage, a movement by which [Utilitarianism] retires to a wholly transcendental

---


standpoint from which all it demands is that the world should be ordered for the best, and that those dispositions and habits of thought should exist in the world which are for the best, leaving entirely open whether those are themselves of a distinctively utilitarian kind or not. If utilitarianism indeed gets to this point, and determines nothing of how thought in the world ought to be conducted, demanding merely that the way in which it is conducted must merely be for the best, then I hold that utilitarianism has disappeared, and that the residual position is not worth calling utilitarianism.\textsuperscript{7}

If Utilitarianism does not survive as a distinctive moral viewpoint, the argument goes, then it is unacceptable as a moral theory. This is a criticism beyond the criticism that Utilitarianism ought not to be accepted as a project. However, Williams does not reach a definite conclusion as to whether Utilitarianism is an unacceptable (incorrect or false) moral theory and leaves this matter "open for discussion".\textsuperscript{8} The same ambiguity translates to the conclusions that my arguments have reached about Self-Interest Theory.

Derek Parfit directly addresses this issue and continues this discussion about whether Utilitarianism and Self-Interest Theory are self-defeating in Part I of \textit{Reasons and Persons}. Since Parfit directly addresses the question of whether Self-Interest Theory is self-defeating, I will drop the strategy of making arguments concerned with Self-Interest Theory only

\textsuperscript{7} Williams, 1973, pp. 134-5.

\textsuperscript{8} Williams, 1973, p. 135
after first discussing their analogues concerning Utilitarianism. I will now move on to some replies to the objection that Self-Interest Theory is self-defeating.
II. Whether Self-Interest Theory is Indirectly Self-Defeating and Not Directly Self-Defeating

The point of Self-Interest Theory is that one ought to maximize the utility accruing to oneself over the entire course of one's lifetime. Parfit makes this claim in the following terms:

(S1) For each person, there is one supremely rational ultimate aim: that his life go, for him, as well as possible.  

According to Parfit, this claim is merely formal. That is, Self-Interest Theory does not specify how in particular an agent ought to behave in order to make his life go as well as possible. More specifically, it does not tell him that one ought to have a disposition to maximize one's satisfaction. As I have been arguing, it might be self-defeating to have as one's over-arching concern the desire to maximize one's self-interest because some other set of dispositions will produce a greater utility for the agent.

Parfit argues that this needn't be considered a problem for Self-Interest Theory. Self-Interest Theory is only concerned with our life going as well as possible. It makes no statement as to what means ought to be used to attain that

---

9 Parfit, 1984, p. 4.
goal. If some collection of desires and dispositions which are unmindful of the concept of utility maximization for oneself happen to be the most effective means of maximizing the utility accruing to oneself, then Self-Interest Theory endorses actions taken to cultivate those desires and dispositions in oneself. This is the indirect pursuit of utility mentioned earlier in the discussion of Williams. However, Parfit has shown that there is no contradiction in Self-Interest Theory urging upon us dispositions which are unmindful of the Utilitarian calculus. For its aim is merely a formal one. Therefore, Self-Interest Theory does not fail in its own terms. In Parfit's terminology, Self-Interest Theory may be "indirectly individually self-defeating". He defines this term such that we can say of some theory T that it is:

\[ \text{indirectly individually self-defeating} \]

when it is true that, if someone tries to achieve his T-given aims, these aims will be, on the whole, worse achieved.\(^{10}\)

Parfit thinks that the fact that Self-Interest Theory is indirectly individually self-defeating would not make it damagingly self-defeating.

Essentially the strategy advanced by this response to the

\(^{10}\) Parfit, p. 5.
charge that Self-Interest Theory is self-defeating is to push Self-Interest Theory out of the role of assisting us in our day to day decisions and to put it in the role of, so to speak, the indirect governor of our lives. It will, thus, guide us in selecting which desires and dispositions we ought to cultivate. These desires and dispositions will, in turn, more directly promote certain actions and outcomes.

The problem, however, is that, as the Williamsonian arguments given in Chapters 1 and 2 point out, the disposition to maximize does take up psychological space. The reason why we acquire a desire or a disposition affects both the success of our attempts to adopt a given disposition and the efficacy of those desires and dispositions to maximize utility.

In relegating Self-Interest Theory to the role of this sort of governor of what desires and dispositions to cultivate, one is essentially talking about character planning in the service of utility maximization. But, as I argued at length in Chapter 2, our motives for modifying our character, whether it be an intentional plan to change one's dispositions or a case in which those changes arise as an unintended byproduct of some other aim, strongly influence whether those character changes are acceptable. Indeed, if that argument succeeds, there is strong reason to think that Self-Interest Theory will be self-defeating even when it enters decision
making at the margin, guiding us in adopting and cultivating various primary desires and dispositions. For one of the points of the argument in Chapter 2 was that if we attempt to manipulate our character merely for the purposes of utility maximization, we are apt to sap the meaning and vitality out of those very projects and commitments which we seek to endorse. In regarding them as a mere means we alienate ourselves from our actions, and from the foundations of those actions in our emotions. Thus, if the argument in Chapter 2 succeeds, Self-Interest Theory seems to have ushered itself not merely to the margins of the scene, to a point of indirect governance of the personality, but rather to have ushered itself from the scene of intentional human action all together.
III. Whether Self-Interest Theory is Self-Effacing and Not Self-Defeating

Introduction

In what follows, I will consider Parfit's argument that Self-Interest Theory may be self-effacing and not self-defeating. What is meant by this is that Self-Interest Theory might tell us to believe some other theory, $y$, in order to maximize utility. Parfit claims that a Self-Interest may tell us what the best thing to do is by telling us to believe some other theory. Therefore, the argument goes, it is not self-defeating. This argument also serves as a reply to the arguments of the first two chapters. For if Self-Interest Theory is self-effacing, then the arguments in those two chapters were based upon a misapplication of Self-Interest Theory, and do not damage it.

I will then reply to the argument that Self-Interest Theory is self-effacing and not self-defeating. My first argument will be based upon what I take to be reasonable constraints upon a theory of rational action. The argument is, basically, that if a theory is a theory of practical rationality it must offer agents guidance as to what they ought to do. Self-Interest theory does not offer agents such guidance if it is totally self-effacing. Therefore, Self-
Interest Theory is not a theory of rationality.

In itself this argument is inconclusive. For Self-Interest Theory might tell us to adopt some other theory, \( y \), and then remove itself from the scene. In that case, Self-Interest Theory does guide the action of the agent to a limited extent and so is invulnerable to the previous objection. I will respond to this move with three counter-arguments. The first of these will be an argument that, if it is true that one's actions that are based upon \( y \)-given reasons are instances of "rational irrationality", then acting on reasons derived from Self-Interest Theory will be an instance of "irrational rationality". That is, one may believe the right theory for the wrong reasons, so that one's actions are, in a sense, irrational. This is what I call the Regress Argument. It is based upon the repeated application of the principle that the reasons why one caused oneself to believe something confers some status of rationality or irrationality upon the actions following from those beliefs.

The next argument is an empirical argument that Self-Interest Theory is self-defeating if it tells us to cause ourselves to believe \( y \). This is because certain features of agents and the concrete circumstances in which they find themselves makes it the case that significant disutility may occur if agents cause themselves to believe some theory of
practical rationality that they believe to be false.

The final argument of this section is an epistemological argument. It is concerned with what grounds we could have for believing Self-Interest Theory. I will argue that if Self-Interest Theory is self-effacing it cannot be justifiable in terms of reflective equilibrium. Reflective equilibrium is not the only means of justifying a theory of rationality, but it is an important one. If we believe that a theory of practical rationality ought to be justified in terms of reflective equilibrium, then we ought not to accept Self-Interest Theory if it is self-effacing.

It should be clear that the first argument of this section against Self-interest Theory as a self-effacing theory works in combination with the last three. If Self-Interest Theory is self-effacing it either provides agents with reasons for action or it does not. If it does not, then the argument in section III B shows why it is not a theory of practical rationality. If Self-Interest Theory does provide an agent with reasons for action, reasons to cause himself to believe some other theory, then the final three arguments give considerations as to why Self-Interest Theory might turn out not to be the true or the best theory. Having discussed the arguments of this section, I will now go on to make them.
A. The Argument That Self-Interest Theory is Self-Effacing and Not Self-Defeating

A theory which is in a position such as that which I described in section II is what Parfit calls "self-effacing".\footnote{Parfit, 1984, Section 9.} It is at this point that Parfit erects a second line of defense to the charge that Self-Interest Theory is self-defeating. He argues that for a theory to be, even entirely, self-effacing is not for it to be damagingly self-defeating.

The basic argument is as follows. For the reasons, discussed in Section I of this chapter it would be bad for us if our behavior were guided by dispositions and motives to maximize self-interest. Similarly, for the reasons given in Section II of this chapter Self-Interest Theory would still be damagingly self-defeating if it told us to attempt to create within ourself any optimal set of reasons and motives in order to indirectly maximize utility. We would be better off if we acted on the basis of some other theory of rationality altogether. In other words, it may be the case that my life will go as well as possible if I believe some theory other than Self-Interest Theory. Since we would best achieve our Self-Interest Theory-given aims if we adopted that theory, we are being most rational, in terms of Self-Interest Theory, if
we adopt that other theory.

None of this implies that Self-Interest Theory is self-defeating. For Self-Interest Theory has directed us to a means of best achieving our Self-Interest Theory-given aims. We simply must adopt a peculiar sort of means of achieving those aims: adopting some other theory of rationality. But this does not mean that this other theory is the true theory, for our only reason for accepting it was that Self-interest Theory endorsed it as a means of making our life go as well as possible. Ultimately, that choice is still grounded in Self-Interest Theory. Parfit Writes:

Suppose that S[elf-Interest Theory] told everyone to cause himself to believe some other theory. S[elf-Interest Theory] would be self-effacing. If we all believed S[elf-Interest Theory], but could also change our beliefs, S[elf-Interest Theory] would remove itself from the scene. It would become a theory that no one believed. But to be self-effacing is not to be self-defeating. It is not the aim of a theory to be believed. If we personify theories, and pretend that they have aims, the aim of a theory is not to be believed, but to be true, or to be the best theory.  

Thus, Self-Interest Theory may remove itself from the scene and still be the true or the best theory. It could tell us to believe some theory which did not give rise to the sorts of alienation discussed in Chapters 1 and 2. If this is so, my

---

12 Parfit, p. 24 (Italics are Parfit's).
third argument against Self-Interest Theory falls flat.

A further point can be drawn from this argument. If Self-Interest Theory is self-effacing, then it also challenges the arguments of the first two chapters. For the Self-Interest Theorist could argue that the sorts of alienation discussed in those chapters arose through a misapplication of Self-Interest Theory. If Self-Interest Theory is self-effacing, it would not enter the psychological space of our decision making in the ways described earlier. The fact that a misapplication of Self-Interest theory makes it appear to be a poor theory of rationality is not a problem with Self-Interest Theory, it is a problem with our use of the theory. Therefore, the arguments of Chapters 1 and 2 do not weigh against Self-Interest Theory if it is self-effacing.

B. Self-Interest Theory Cannot Be Completely Self-Effacing And Still Be A Theory of Practical Rationality.

Suppose that Self-Interest Theory is totally self-effacing. That means that Self-Interest Theory will tell us to believe some other theory, \( y \). But then it is \( y \) that gives
us those reasons for action that make our life go as well as possible. y is the theory which, if we believe it, will make our lives go as well as possible.

Parfit tells us that this does not mean that y is the true or the best theory. But consider the following: when we still act according to our Self-Interest Theory-given reasons our lives do not go as well as possible; when we adopt y our y-given reasons give us motivation to act in a way that makes our life go as well as possible. This latter claim is true because once we adopt y we are motivated by it, not because it is endorsed by Self-Interest Theory, but because we accept it as the true theory. We believe y. If we did not, we would have the same jaded perspective on the motivations produced by y as Bill, in the example in Chapter 2, did to potential motivations whose only source of value to him lay in their potential to produce personal utility. Therefore, Self-Interest Theory does not provide a system of non-self-defeating reasons for action for agents. Stated in other terms, Self-Interest Theory would never provide a system of reasons for action for an ideally rational agent.

If a theory of practical rationality only offers an agent a system of reasons for action which are self-defeating, then it cannot be the true or the best theory of practical rationality. If a theory offers an agent no reasons for
action at all, then it is not a theory of practical rationality. Theories of practical rationality are concerned with what intentional actions an agent ought to take. The idea of intentional action presupposes that the agent has reasons for action. Thus, though they may have other features, theories of practical rationality must at least be concerned with agents both having and acting upon the best reasons for action. An agent cannot act upon reasons for action if he does not have them as a part of his conceptual scheme. Therefore, if a theory provides an agent with no reasons for action, it is not a theory of practical rationality at all. If a theory provides an agent with only a self-defeating system of reasons for action it cannot be the true or the best theory.

If Self-Interest Theory does not provide reasons for action, there is some question as to how it gains its intuitive force as a theory. One answer to this question could be that the maximizing notion which is a component of Self-Interest Theory is better understood as a component of theories of the good for agents. This idea fits well with the discussion in Chapter 1 of how Self-Interest Theory requires the agent to stand outside of his projects and commitments when determining how to maximize. Thus, in looking at one's own life and all the cares and concerns that make it up from the standpoint of Self-Interest Theory, one is looking at it
as he might look at the life of any agent that he wanted to see do well. He regards the projects he sees from this perspective not as projects to which he is immediately attached, but as projects of which he seeks the maximally harmonious satisfaction.

This interpretation also fits well with the fact that for many concrete goods the concept of maximization isn't immediately applicable. Gaining increasing quantities of some goods will, in some cases, cause the agent to value each next unit of that good less. This is the law of diminishing marginal utility. In other cases, gaining increasing quantities of a good will produce an aversion to that good.\textsuperscript{13} More is not always better where concrete ends are concerned. In speaking of maximizing happiness or preference satisfaction one is seeking an underlying unit of value which is not subject to the vicissitudes of the various concrete rewards. But here one has already taken up a position outside of particular aims and interests. One is seeking a universal currency in terms of which goods for agents are commensurable. Once the good for agents is defined in this way, the notion that "more is better" follows as a sort of corollary. Thus, the intuition that lies behind Self-Interest Theory could be that what is best for an agent (any agent) is that his life go as well as possible, where "better" is tied to some maximizing

\textsuperscript{13} Ainslie, Chapter and Section 1.3.3.
notion.

As we have seen, the notion of one's life going as well as possible could be cashed out in terms of maximization of happiness or of preference satisfaction. This conception could either be a theory of the good for agents, or a component of some such theory. If one is seeking the good of some agent, he may try to facilitate the agent's aims such that the agent's life goes as well as possible. The intuitions behind Self-Interest Theory, then, may support a theory of the good for agents rather than a theory of practical rationality for agents. It may be that the aims of the former cannot be directly translated into the agents reasons for action. It has been a constant theme of this paper to suggest that the first person perspective on one's life and the external or objective perspective on one's life may not be directly commensurable.14

I do not believe that I have offered a conclusive argument for the claim that the intuitions which lend their support to Self-Interest Theory are better understood as intuitions supporting a theory of the good for agents rather than

---

14 Thomas Nagel has done a great deal of important work concerning this issue. See for instance Nagel, 1970 and Nagel, 1986. A full response to Nagel's criticisms of positions like that for which I argue and that for which Williams argues would be a very major undertaking. I do not specifically respond to Nagel at any length within this thesis.
than a theory of practical rationality. Indeed, the argument has only been suggestive. However, it does offer a possible means of accounting for the powerful intuitions that there is something importantly right about Self-Interest Theory\textsuperscript{15}, while at the same time claiming that it is either not the true or the best theory of practical rationality, or not a theory of practical rationality at all.

\textsuperscript{15} For an example of how a theory based upon a conception of agents as rational maximizers can have very useful descriptive and predictive, as well as normative, applications see Becker. Becker applies the economic model of rationality to a number of areas of human behavior including law, politics, marriage and social interactions which were traditionally considered non-market activities. It is interesting to note that Becker concludes that even if individual agents or units of production do not maximize, the effects of "irrationality" may cancel out such that the assumption of maximization is still fruitful for in describing overall market behavior. See, for instance, Becker Chapter 8. Becker writes:

"...little attention has been paid to the distinction between group or market and individual responses. This distinction is unnecessary in traditional theories of rational behavior because a market's response is usually simply a macro-version of an individual's response. A group of irrational units would, however, respond more smoothly and rationally than a single unit would, and undue concentration on the individual level can easily lead to an overestimate of the degree of irrationality at the market level....the negative slope of market demand curves or the positive slope of market supply curves, is equally consistent with individual irrationality and cannot distinguish between [it and individual rationality]." (p. 167-168)

Thus, the model of the self-interested maximizer is useful for understanding the behavior of groups, even if it does not precisely mirror how individuals do in fact behave at the individual level, or, indeed, ought to act at the individual level.
There is a reply to this argument that Self-Interest Theory is not a theory of practical rationality at all. The Self-Interest Theorist's first reply could be that Self-Interest Theory could form a part of the system of reasons of a fully rational agent in that Self-Interest Theory provides the agent with reasons to believe \( y \). In giving the agent these reasons it helps him to make his life go as well as is possible. Therefore, Self-Interest Theory does prove to be part of the system of reasons of the ideally rational agent and so escapes the charge of being either self-defeating or not a theory of practical rationality at all.

To this we might reply that Self-Interest Theory might form a component of a non-self-defeating system of reasons in telling us to believe \( y \), but at a very high cost. For if Self-Interest Theory tells us what to believe, namely to believe \( y \), it is not serving as a theory of practical rationality. For a theory which tells us what we ought to believe, rather than what we ought to do, is not a theory of practical rationality, it is a theory of theoretical rationality. Therefore, Self-Interest Theory survives the charge that it never forms a part of a system of reasons that is not self-defeating by ceasing to be a theory of rational action at all. For, as it was already argued, once we adopt \( y \) we act on our \( y \)-given aims and reasons. Before we adopt \( y \), we will be defeating our Self-Interest Theory given aims by
attempting to maximize our self-interest. Thus, Self-Interest Theory can never form a system of reasons for agents which is not self-defeating and so, as a theory of practical rationality for agents, it is self-defeating.

A natural line of response to this argument is to say that Self-Interest Theory does not merely tell us to believe \( y \), it tells us to cause ourselves to believe \( y \). Causing ourselves to believe something is a kind of action. What is more, it is not self-defeating action in that the indirect result of these actions is that we cause our lives to go as well as possible. Thus, Self-Interest Theory survives as a theory of practical rationality. This is a strong response, and must be considered further. I will now go on to consider three counter-arguments in response to this argument.

C. The Regress Argument Against Self-Effacing Theories of Practical Rationality

So far we have seen that Self-Interest Theory might give an agent reasons for action in telling him to cause himself to believe \( y \), and that these reasons might not be self-defeating.
In discussing these issues Parfit presents several examples which are intended to show how it might be better for us, that is, a more rational course of action, if we caused ourselves to behave irrationally. For by behaving irrationally one can, in certain circumstances, better achieve ones aims. This is what Parfit calls "rational irrationality". If we think of our acting on our belief in y as a case of rational irrationality, we might believe that the fact that we had reason to cause ourself to be irrational confers the status of rationality upon the whole episode.

However, this line of argument raises a problem when it is extended from particular instances of action to our motivational system of practical rationality as a whole. For if acting on y is rational irrationality because my belief in y was caused by my actions which were undertaken because they were directed by my truly rational aims, then it is unclear why this move cannot be pushed back further still. In other words, if the activities which comprise my life can be dubbed "rational", even though I believe y, merely because my belief in that theory was caused by my Self-Interest Theory-given motivations, why cannot the same move be made with regard to my motivations for adopting Self-Interest Theory?

Suppose that an agent is not born believing Self-Interest

---

16 Parfit, Part I Sections 5 and 6.
Theory. If this is so, the agent would initially be irrational in terms of practical rationality. Hence, whatever moves the agent takes to cause himself to believe, or maintain a belief in, Self-Interest Theory would be irrational in that they would be motivated by reasons which are not a part of the best theory. Therefore, in acting on in Self-Interest Theory the agent is being irrationally rational, and in acting on y he is being irrationally rationally irrational.

If the foregoing conclusion is right, then not only my actions taken in accord with y, but my actions taken in accord with any other theory are ultimately founded on irrationality. Therefore, there is no such thing as rational action per se. But this is an absurd conclusion and must be rejected. Therefore, we must not look at acting on y as rational irrationality when Self-Interest Theory encourages total conversion to y. We must look at actions in accord with y simply as rational actions. If this is so, then Self-Interest Theory is self-defeating. For y directs us to take actions we ought rationally to take, not Self-Interest Theory. Self-Interest Theory has no place in the optimum set of reasons for

---

17 I will not argue for this assumption. It is obvious that children aren't effective maximizers, but it isn't clear whether this is because they are not Self-Interest Theorists or because they apply Self-Interest Theory poorly due to their limited cognitive abilities. I will not attempt to tackle this thorny question of interpretation. If the reader does not accept the assumption that children are not naturally Self-Interest Theorists, then he will not find the argument convincing.
action which makes an agent's life go as well as possible. It is, therefore, self-defeating.

There is the possibility that the agent could be better understood as choosing to adopt Self-Interest Theory on the basis of some meta-level theory. This would be a theory about which decision procedure one should adopt for the purposes of deciding how to act. It would be a theory of theory choice for deciding among competing theories of practical rationality. If this is so, I have no grounds for saying that actions taken to cause oneself to believe Self-Interest Theory were irrational because they were guided by non-Self-Interest Theory-given reasons. Self-Interest theory would not provide the criteria of rationality at this level. Therefore, the regress argument would fail.

In response to this objection I make the following claim: theory of practical rationality--the theory of how one ought to act, and theory of theoretical rationality--the theory of what one ought to believe, are exhaustive of the categories of human reason. At the highest level, so to speak, one can still only decide what to believe or how to act. One is, therefore, operating either in terms of theoretical rationality or practical rationality. If Self-Interest Theory is the best theory of practical rationality, then I take it that Self-Interest Theory provides the criteria for rational
action "all the way up", so to speak.

The Regress Argument is itself an attempt to criticize a move which places Self-Interest Theory at the meta-level. For this objection to the Regress Argument to succeed there would have to be at least three tiers of rationality for actions. There would be a theory of theory choice, \( t \), governing the choice of a formal theory of practical rationality, Self-Interest Theory, which governs the choice of some theory which best achieves our Self-Interest Theory-given aims, \( y \). This seems implausible if \( t \) is not, in fact, simply a theory of theoretical rationality. If \( t \) is a theory of theoretical rationality, other problems emerge.

Suppose the Self-Interest Theorist claimed that the ideally rational agent needn't cause himself to believe Self-Interest Theory in the way that I have described it thus far. Self-Interest Theory could be believed because the agent has reasons to believe it. Thus, belief in Self-Interest Theory could be grounded in considerations of theoretical, rather than practical, rationality. Belief in \( y \), on the other hand, was caused by actions taken on the basis of practical rationality.

Several replies can be made to this argument. First of all, it offers a rather optimistic picture of how agents come
to have and maintain beliefs. For often the reasons or causes which lead agents to acquire or maintain beliefs is very different from whatever reasons they may have which make those beliefs justified. Still, we are talking of ideally rational agents. Granted the distinction between theoretical and practical rationality, the position that belief in Self-Interest Theory can be sustained for the right justificatory reasons can be maintained. However, I will show in Section III E of this chapter that it may be difficult for an agent to justify his belief in Self-Interest Theory if it is self-effacing.

One could also argue that the proper theory of theoretical rationality is that one ought to believe those propositions which maximize rationality. In that case, unless one maintained the position that people are born maximizers, they would acquire a belief in Self-Interest Theory for non-rational reasons. If that is so, then the regress argument could be employed. If acting on $y$ is rational irrationality, then acting on Self-Interest Theory is irrational rationality. The absurd conclusion that there is no rational action simpliciter follows, and so we must reject the argument that Self-Interest Theory is self-effacing and not self-defeating.

Suppose however that we allow for a firmer distinction between theoretical rationality and practical rationality. If
Self-Interest Theory tells us to cause ourselves to believe $y$ it is telling us to believe a false theory about what one ought to do. As a result, the agent will also be mistaken as to all the reasons he gives as to why a given course of action is rational or not. The entire scheme upon the basis of which he governs his action and most of his beliefs about that scheme will be false. He will be globally deluded with regard to knowledge about what he ought to. When he believes that some act, $A_1$, is the rational action, and it is, he still believes it is such for the wrong reasons. It is a sort of happy coincidence that his belief is true. He believes that $A_1$ is the right action for the wrong reasons, he believes it for $y$-given reasons. Had the agent not caused himself to believe $y$, he would not lack this sort of knowledge.

Assuming, then, that theoretical and practical rationality are distinct concerns of the agent, the question arises as to why the agent ought to be willing to abandon one for the sake of the other. That is, it is not clear that the agent ought to cause himself to be egregiously theoretically irrational by causing himself to believe $y$ in order that he maximize utility. This is a difficult issue to pursue further without developing more fully a theory of theoretical rationality, and a more general theory of what relationship the two sorts of rationality bear to one another. I will not be able to carry out that project within the limits of this
This is not a conclusive objection to Self-Interest Theory. However, it does seem to weigh heavily against Self-Interest Theory, making it less plausible as an account of practical rationality. For if the forgoing argument is correct, the two sorts of rationality of which an agent is capable would be at odds with one another if Self-Interest Theory is self-effacing. This seems particularly problematic in that Self-Interest Theory evaded the initial regress argument by presupposing that the agent under consideration was not only ideally practically rational, but was also ideally theoretically rational, and so had adopted Self-Interest Theory for the right theoretical reasons.

D. The Empirical Argument That Self-Interest Theory is Self-Defeating Even If It Is Self-Effacing

An argument can be made based upon empirical considerations that Self-Interest theory is self-defeating. This argument centers on considerations about the characteristics of agents and the contexts in which they
operate. I will argue that if Self-Interest Theory tells agents to cause themselves to believe some other theory, then it is self-defeating.

In the simplest case the agent, through chance and circumstance, believes $y$, the theory that will make his life go as well as possible from the start. Perhaps he is raised in a culture which educates its children in the philosophy of $y$. Here Self-Interest Theory would tell us to do nothing. Self-Interest Theory would therefore not guide any action and would not serve as a part of the agent's system of reasons for action. It is therefore open to the objection discussed in Section III B.

Should an agent in such a society begin to lose his belief in $y$, Self-Interest Theory would tell him to maintain his belief in $y$. Similarly, if the agent does not already believe $y$, Self-Interest Theory would tell him to cause himself to believe $y$. Here the agent has entered into a tricky situation. If he causes himself to believe $y$ in a less than total way, he will have, somewhere in his mind, the knowledge that he had reason to cause himself to believe $y$, and reason to cause himself to believe that $y$ is the best theory, but no reason to believe that $y$ is true. Thus, he had reason to cause himself to engage in massive self-deception. Knowledge that the system of one's beliefs and reasons that
guide his life is based on massive self-deception or bad faith is likely to produce serious negative psychological consequences. Even if this knowledge is largely suppressed it is likely to have its effect in nagging feeling that one is selling out, is living in bad faith, or is simply a fraud. It is hard to see how such feelings could be consonant with having one's life go as well as possible.

To escape this difficulty, the individual may take action such that he, in effect, has a total conversion to y. This may solve the problem of the malaise associated with living in bad faith. However, several new problems crop up.

The first problem is that, having thrown away his rudder, so to speak, the agent is cut adrift on y and must go wherever that theory takes him. The agent might move to a new location, changes might occur in his social group or family or culture, his personality may change with age or experience, a new technology may affect the benefits or drawbacks of a given course of action or the choices and opportunities he may face, or the agent may simply be thrust into situations significantly dissimilar to those in which he found himself in the past. Under new circumstances, y may no longer be the theory which, if believed, would best promote the agent's Self-Interest Theory-given aims. In this case, Self-Interest theory has directed the agent to take actions, the indirect
result of which, is to cause an agent to worse achieve his Self-Interest Theory-given aims than if he believed some other theory.

It is true that, over time, an agent may come to change part of his theory y, or to adopt some wholly new theory. But this will be undertaken because y, or some other theory, directed him to do so. After his total conversion Self-Interest Theory will not guide him as to which theory he ought to believe. Therefore, given the various theories that the agent could have, it is unlikely that the new theory that y picks out will happen to be the one that best promotes self-interest.

It could be replied to this argument that Self-Interest Theory would only lead the agent astray in such situations if y was not the best theory. Thus, the difficulty is not with Self-Interest Theory, but with the ignorance of the agent. If the agent had perfect information and chose the theory that best promoted his Self-Interest Theory-given aims, then that theory would, in fact, best promote those aims. The agent's error is not the fault of Self-Interest Theory.

To this I would respond that we are here concerned with a theory of rationality for finite agents, not for gods. So far I have already brought into the argument all manner of
contingent features of agents, such as under what circumstances agents experience alienation. These considerations were brought in order to show that Self-Interest Theory is a self-defeating theory of rationality for agents. To allow for agents to have the sort of vast cognitive abilities that would allow them to predict the sort of changes described above, as occur in ordinary human situations, and how the agent would fare in each of these possible situations were he to adopt each of the possible theories of practical rationality, would be an arbitrary and unrealistic extension of the cognitive powers of actual finite agents.

Even if agents could determine the consequences of adopting each of the various theories, the expenditure of time, energy, and resources in making such a determination would be great. There is good reason to think that the expenditure would be great enough such that even if the agent picks the right theory it would not compensate for the costs of decision making. The agent would have been better off had he believed some other theory and never believed Self-Interest Theory. He would have been better off, in self-interested terms, had Self-Interest Theory never entered his system of reasons for action. Thus, Self-Interest Theory is either not a candidate for a theory of practical rationality, or it is a self-defeating theory. We are looking for a theory of
rationality which is not self-defeating for ordinary human agents.

It might, nonetheless, be argued that in cases where the agent had sufficient foresight and a sufficiently stable environment he could, in fact, determine which theory would best promote his Self-Interest Theory-given aims. He could then effect a conversion to that theory. If we assume that the agent does have sufficient cognitive skills and abilities to pick out the best theory in terms of his Self-Interest Theory-given aims, there nonetheless arise several other sorts of problems in causing himself to believe that theory.

If accepting theory \( y \) involves adopting some shared values, teachings or beliefs, the agent may encounter further difficulties. For if he seeks to adopt the values or ways of life of some settled cultural, ethical, religious, spiritual or philosophical community, the members of such a community may resist or reject him because he has come to share their beliefs or values not because he has good reason to believe them to be true, or had some religious experience of conversion, but rather because he has ulterior self-interested motivations to cause himself to believe them. The agent is apt to be considered dishonest, ingenuous or simply as one living in bad faith. If the community does not have such scruples it seems unlikely that the way of life which they
promote would tend to promote the aim of having one's life go as well as possible. For the members would have to admit that they practice a way of life that admitted of massive self-deception. This fact would be brought to their attention whenever they admitted new members to the community who were motivated by purely self-interested reasons. The tension or malaise that such a situation would produce would almost certainly ensure that the form of life practiced by that community did not best secure the happiness of agents. Thus, there are strong empirical reasons to believe that causing oneself to believe in $y$ would not best promote one's Self-Interest Theory-given aims. If this is so, Self-Interest Theory is a theory which cannot be a part of an agent's systems of reasons without being self-defeating.

One is left, therefore, with the prospect that Self-Interest Theory provides non-self-defeating reasons for action only when it gives us reason to act so as to cause ourselves to believe some other theory, $y$, by means of some total conversion, under circumstances that are sufficiently stable, and when the conversion technique does not involve intimate and sustained social relations. It is unclear what means of acting would, in practical terms, be left open for the agent. In order to remove all significant traces of belief in Self-Interest Theory the agent would probably have to engage in something like brainwashing, either through the help of some
other agent or agents, or by some rather radical means of acting on oneself. Such agents as would help one cause oneself to believe some theory that one took to be false would be unlikely to be agents who participated in the rational or ethical tradition which one wished to enter for the reasons previously discussed. They would have some motivation for offering their assistance other than sharing their learning or values. Such agents would be likely to be persons who were motivated by some external reward (such as money) for their assistance to the agent in his conversion.

Presumably causing oneself to adopt the new theory would involve some extreme regimen of operant conditioning. At this point, it, again, becomes clear to what extent Self-Interest Theory licenses us to take an external perspective on ourselves so that we operate upon ourselves in order to cause ourselves to believe some theory which we know to be false, and allow that theory to guide our lives. This seems a rather grotesque sort of self-manipulation. In any event, there is much room for doubt that such conditioning could provide the agent with a sufficiently complex belief system to best achieve his Self-Interest Theory-given aims. Therefore, even within the limiting conditions described at the beginning of the last paragraph, it is not clear how Self-Interest Theory could provide reasons for action without being self-defeating.
E. Problems Grounding Self-Interest Theory If It Is Self-Effacing

There is another argument against Self-Interest Theory if it is a self-effacing theory. This argument raises the question of what grounds we could have for believing Self-Interest Theory to be the true or the best theory if it is self-effacing. I will argue that, if Self-Interest Theory were self-effacing, we would not be able to justify that theory in terms of reflective equilibrium. Thus, an important grounds for believing Self-Interest Theory to be true would be lost.

In trying to discern what evidence we could have for believing Self-Interest Theory to be the true or the best theory we ought to reject the notion that evidence could be found in terms of observations of human behavior. By "behavior" I mean bodily activities as they are understood without reference to the aims or ideals of the actors. Any such observations could, at best, only tell us what agents happen to do, not what they ought to do.

There is a second problem with attempting to support a theory of normative practical rationality in terms of human behavior. Any collection of human behavior can be interpreted in a variety of ways either as intentional or unintentional
action. In After Virtue, Alasdair MacIntyre writes:

It is a conceptual commonplace, both for philosophers and for ordinary agents, that one and the same segment of human behavior may be characterized in a number of different ways. To the question 'What is he doing?' the answers may with equal truth and appropriateness be 'Digging', 'Gardening', 'Taking Exercise', 'Preparing for winter' or 'Pleasing his wife'. Some of these answers will characterize the agent's intentions, other unintended consequences of his actions, and of these unintended consequences some may be such that the agent is aware of them and others not.\(^1^8\)

He concludes that "We cannot...characterize behavior independently of intentions".\(^1^9\) Without arguing the point further, I will claim that intentions cannot simply be read off physical movements. Intentional action requires that the agent have some intention behind his action. This intention can only be understood once one has grasped the sorts of aims and purposes the agent might have.

Knowledge of an agent's intention might be gained either by inquiring about it with a particular agent, or by making an interpretation based upon relevant knowledge regarding similarly placed agents, what one would do in similar circumstances, and so forth. The point is that somewhere along the line one has to use either reflective knowledge of oneself, or use knowledge gained through communication with

\(^{18}\) MacIntyre, p. 206.

\(^{19}\) MacIntyre, p. 206.
other agents, in order to make sense of an intention behind an action. Therefore, no evidence for a theory of rational action, including Self-Interest Theory, could be gained merely on the basis of data about physical movements.

This brings us to the second point. Often theories about what agents ought to do are grounded in what is called "reflective equilibrium".\(^{20}\) This is achieved by obtaining a balance between one's basic principles and his considered judgement of cases. Suppose Self-Interest Theory were self-effacing. If this were so, Self-Interest Theory could not be grounded in reflective equilibrium.

If Self-Interest Theory were self-effacing it would be believed, at best, only briefly by agents. For it would instruct them to come to believe some other theory, \(y\), which helped their life go as well as possible. If it did not do this, or if the agent waited a long time to cause himself to believe the new theory, Self-Interest Theory would be self-defeating. Some agents would never come to believe Self-Interest Theory at all. Given these facts, one would not expect to find support for Self-Interest Theory in the intuitions of agents. Those who had never come to believe the theory would not have maximization of their individual utility as their principle of rationality. Those who came to believe

\(^{20}\) Rawls, Chapter I, Section 8.
Self-Interest Theory would, if they behaved rationally, adopt some other theory, \( y \). Therefore, they would not have as their central principle of rationality the aim that one's life go as well as possible. Thus, one would not expect to find support for Self-Interest Theory in reflective equilibrium between the considered judgements of agents and their principles of rational action. For though one might have it among his aims that, in some respects, it is better that his life go as well as possible in terms of utility maximization, this would not form the central principle of rationality for him once he adopted \( y \). In addition, many of his considered judgements might contradict the central aim of Self-Interest Theory. After all, \( y \) is, strictly speaking, irrational from the standpoint of Self-Interest Theory.

However, some agents might believe Self-Interest Theory and not adopt some other theory, \( y \). They might have intuitions that Self-Interest Theory is the true theory. Supposing Self-Interest Theory to be self-effacing, they might fail to adopt theory \( y \) for one of three reasons. First, they may lack sufficient information to realize that Self-Interest Theory is self-effacing. Secondly, they may be incapable of causing themselves to believe some other theory. Thirdly, they may choose to be practically irrational rather than theoretically irrational.
If Self-Interest Theory is self-effacing for reasons like those given in Chapters 1 and 2, then it will become obvious to most agents over the course of their life that it is self-effacing, or in some sense that they ought not to believe it. Whether Self-Interest Theory is the true theory or not, I believe that many people do, at some level, have the intuitions that if they lived according to Self-Interest Theory they would suffer the sorts of problems discussed in this thesis. Theoretical sophistication needn't be that which guides the agent to another theory. The unpleasantness of one's life in following Self-Interest Theory will tend to produce that effect. I believe it to be the case that, in fact, few people are genuinely guided by something like Self-Interest Theory in their daily lives. If Chapters 1 and 2 are correct, a reconstruction of people's intuitions about what they ought to do would support this contention. Thus, we would expect a small minority of agents to be in this position.

Suppose agents believed Self-Interest Theory was the best theory but were unable to cause themselves to believe some other theory. If this problem were widespread, Self-Interest Theory would be self-defeating. For if Self-Interest Theory cannot be self-effacing and we followed our Self-Interest theory-given aims our lives would not go as well as possible. Nonetheless, there could be a small class of persons for whom,
for one reason or another, this was true. For them, Self-Interest Theory would be self-defeating, though they might still believe it to be the true theory.

Finally, there might be a class of persons for whom it might be preferable to believe Self-Interest theory rather than $y$. They might find it to preferable to be practically irrational rather than theoretically irrational. They might say that it is more important to believe the right theory than to have a happy life. But then these people have made a value judgement which contradicts the central claim of Self-Interest Theory. Therefore, they are not actually Self-Interest Theorists. They do not believe that Self-Interest Theory really tells them what they ought rationally to do.

We are thus left with two classes of persons who might have intuitions that Self-Interest Theory is the true or the best theory, but who do not cause themselves to believe some other theory $y$. These are: those who do not yet realize that Self-Interest Theory is self-effacing, and those who cannot cause themselves to believe some other theory. If Self-Interest Theory is self-effacing, these intuitions are unlikely to be held by many persons, or the community at large. It is, therefore, unlikely that Self-Interest Theory could be grounded in reflective equilibrium.

There may be other means of justifying belief in Self-
Interest Theory other than through reflective equilibrium, so this argument is inconclusive. One might, for instance, have a priori reasons for believing Self-Interest Theory to be the true theory even though it is self-effacing. Depending upon one's moral epistemology, perhaps Self-Interest Theory could still be defended. In this thesis I will not investigate the various means of justifying moral claims in order to determine whether or not Self-Interest Theory could be justified in terms of these theories even if it were self-effacing. I have merely argued that an important means of justifying moral theories, reflective equilibrium, is unavailable to Self-Interest Theory if it is self-effacing.
CONCLUSION

If Self-Interest theory is self-effacing and not self-defeating, then this would weigh against the arguments given in Chapters 1 and 2. In those arguments I showed that adherence to Self-Interest Theory causes the agent to be alienated from his actions, from his commitments and from his character. In Chapter 1 I argued this on general grounds, based on argument which paralleled Williams' argument against Utilitarianism. The argument centered around the claim that the standpoint of rational assessment for Self-Interest Theory is a standpoint external to the agent's projects and commitments. Adopting such a standpoint, I argued, alienated one from crucial elements of his life.

In Chapter 2 I examined an example of what is typically taken to be a paradigmatic technique of self-interested maximization, namely precommitment, and showed that it could best be interpreted if we assume that the agent operated on the basis of motives other than that of utility maximization. What is more, he is better off, in self-interested terms, acting on these other motives. If the agent were motivated merely by utility maximization, he would be alienated from his own character as agent. I then argued that this result could be generalized to most cases of precommitment, and that this lent support to the general analysis of rationality and
motivation given in Chapter 1. Together, these arguments make Self-Interest Theory appear to be an unacceptable theory of practical rationality.

However, if Self-Interest Theory is self-effacing it could tell the agent to believe some other theory of practical rationality which would not produce alienation in the ways discussed in the first two chapters. In that case, Self-Interest Theory could escape these criticisms. Thus, much hangs on the question of whether Self-Interest Theory could be self-effacing and still be a proper theory of rationality.

Several of the arguments in Chapter 3 were directed toward arguing that a theory of practical rationality could not be self-effacing and still be the true or best theory of rationality. These arguments were introduced after first introducing the claim that Self-Interest Theory could be self-effacing and not self-defeating in response to the charge that Self-Interest Theory is, in fact, self-defeating. The structure of those arguments, in more detail, is as follows.

In Chapter 3 I first argued that Self-Interest is self-defeating based upon the conclusions of Chapters 1 and 2 with respect to alienation. This argument was followed by a reply that Self-Interest Theory is only indirectly self-defeating and not damagingly self-defeating. I then argued against this
reply that Self-Interest Theory was more than indirectly self-defeating, and was, at least, self-effacing. This is because the results of Chapter 2 demonstrate that similar problems of alienation arise with the indirect pursuit of utility as arose with more direct strategies. I then considered the reply to that argument that Self-Interest Theory is self-effacing and not self-defeating.

To the last reply, that Self-Interest Theory is self-effacing and not self-defeating, I gave four counter-arguments. The first such argument was that Self-Interest theory could not be a theory of practical rationality for agents if it was wholly self-effacing, and that the intuitions supporting it might be better understood as supporting either a theory of the good for agents or a component of such a theory. I then considered whether Self-Interest Theory could be a self-effacing, but not self-defeating, theory of practical rationality if it told us to cause ourselves to believe some other theory. This would help Self-Interest Theory escape the first objection against it as a self-effacing theory. I then advanced three arguments against this move. The first argument was a regress argument. This argument attempted to show that the notion that it could be rational to cause ourselves to believe a false theory of practical reason led to an absurd conclusion, and so ought to be rejected. I then went on to argue, based upon empirical
considerations, that Self-Interest Theory would be directly self-defeating if it told us to cause ourselves to believe some other theory. Finally, I advanced an epistemological argument. This was the argument that if Self-Interest Theory were Self-Effacing it could not be justified in terms of reflective equilibrium. Reflective equilibrium is an important means of justifying theories of rationality and morality. This conclusion weighs against Self-Interest Theory.

I do not believe that the last three arguments are conclusive. The regress argument requires either some assumptions about how agents acquire and maintain belief, or a more fully articulated conception of theoretical rationality, in order to be conclusive. The empirical argument makes empirical claims that would have to be supported by psychological and sociological research in order to be conclusive. The epistemological argument only attacks one possible means of justifying Self-Interest Theory. Nonetheless, these arguments do show that there is reason to believe that Self-Interest Theory may not be the true or the best theory if it is self-effacing. The arguments make that theory seem less plausible. Thus, there is good reason to believe that we should not accept Self-Interest Theory in our live as the theory of practical rationality upon the basis of which we act, and good reason to doubt that it is the true or
the best theory.
BIBLIOGRAPHY OF SOURCES CITED


   (a) "Persons, Character and Morality"
   (b) "Moral Luck"
   (c) "Utilitarianism and Moral Self-Indulgence"