Rationality and Self-Conflict

by

Jacob Goldin
Class of 2007

A thesis submitted to the faculty of Wesleyan University in partial fulfillment of the requirements for the Degree of Bachelor of Arts with Departmental Honors in Government and Economics.

Middletown, Connecticut  April, 2007
**CONTENTS**

Acknowledgements

**Introduction**  

**ONE**  
A Model of Rationality  
Basic Concepts  
Conditions on Rational Preferences  
Self-Interest and the Content of Preferences  
Higher Order Preferences  
Rationality and Morality  
Revealed Preferences  

**TWO**  
Making Decisions about the Future  
Dynamic Conflict  
Prudence and Personal Identity  
Conditional Priorities  
Discount Rates and Pure Time Preference  

**THREE**  
Weakness of Will  
Can Weakness of Will Exist?  
Willpower  

**FOUR**  
Self Control  
Strategies of Self Control  
Sticking to a Resolution  
Credibility and Precommitment  

**FIVE**  
Rational Irrationality  
Non-Optimizing Dispositions  
Strategic Dispositions  
Morals by Agreement  

**Conclusion**  
Decision-Making and Social Forces  
Choice and Well-Being  

Bibliography
I am grateful to the many people who helped make this thesis fun to write. Thanks to my friends from home and my friends from school – the ones that are double-counted deserve twice as much thanks. I am especially indebted to my housemates: Albert, Evan, Micah, and Steve, for their willingness to discuss the absurd thought experiments I confronted them with throughout the year. Thanks to my parents, my sister, and my grandparents for being great people in my life, and for teaching me how important it is to have an optimistic attitude. I am also grateful to Debi, whose sense of humor and energy for life have helped make these past four years so enjoyable.

I would be remiss if I did not acknowledge Michael Hanson, from whom I learned more than any other teacher at this school. I also owe a huge debt of gratitude to Gil Skillman for providing me with the intellectual tools that inform this thesis, as well as for commenting on sections of it. Finally, Don Moon’s encouragement is what prompted me to take on this project in the first place, and his enthusiasm and good humor have been refreshing throughout the entire process. The ideas presented here have benefited enormously from his thoughtful guidance.

The Davenport Committee supported the research project that turned into this thesis, and I gratefully acknowledge their generosity.
Rationality has fallen on hard times. The dominant approach to modeling decision-making in the social sciences has always been controversial, but in recent years, the attacks on rationality have cut especially deep because they have come from inside the discipline most closely associated with the approach: economics itself. As behavioral economics has emerged as a respected subfield, its practitioners have obtained results that cast doubt on many of the core assumptions that inform traditional models of decision-making. Although these findings have received a great deal of attention within the economics literature, they have been less surprising to those outside the discipline. The most common reaction seems to be that economists are finally discovering what the rest of us have always known: that real people’s behavior does not match the understanding of rationality embodied in homo economicus.

Even more striking is the fact that many of these alleged deviations from rationality do not trouble the people who commit them. For example, the laboratory finding that most people are quite willing to sacrifice real material gains out of concern for others’ well-being does not fit well with the picture of rationality in most economic models of decision-making – a picture that is central to many of the discipline’s proudest results, such as the Arrow and Debreu theories of general equilibrium. But for most of us, such findings are cause for comfort rather than concern; people who consider the values embodied in homo economicus to be overly narrow and selfish do not view it as a mistake when they act on values other than
those. Similarly, the picture of rational agents as constantly optimizing to make the best possible decision in every circumstance is a way of behaving that most people find inaccurate, but also inconsistent with how they would like to lead their lives. These disparities – between how people assess the quality of their own decisions and what theorists consider rational – point towards a more fundamental problem with the standard model of rationality: that the behavior it prescribes is ultimately undesirable. In other words, people see rationality as an ideal, but not as an ideal that is worth pursuing.

Many of the criticisms leveled against this picture of rationality are valid. The conception of rationality that informs most decision-making models in the social sciences depicts people as acting in ways that, even if accurate, would not be desirable because of the narrow restrictions imposed on their values and the important features of choice situations that get ignored. The flaws in this conception of rationality, while often defended as convenient for descriptive purposes, are crippling for normative applications. We can identify at least four reasons why such normative inadequacies are cause for concern.

Most basically, a normatively compelling account of rationality is important for prescriptive purposes. Theories of rationality prescribe how decision-makers should act in order to best achieve their goals. In doing so, they shed light on what types of strategies can be productive for decision-makers to adopt, and in what circumstances. Decisions that do not meet the standard of rationality cannot be defended on the basis of the decision-maker’s values, and are thus subject to criticism.
at a basic level. A conception of rationality that is normatively flawed – that does not prescribe desirable behavior – will fail at these basic tasks.

Second, a flawed conception of rationality undermines other normative theories in which rationality plays a role. For example, models of rationality occupy prominent positions in the works of political theorists ranging from Thomas Hobbes to John Rawls.¹ Similarly, any normative theory that measures one’s freedom as more than the number of available options must rely on a theory of rationality to determine which choices people have reason to value.² If the model of rationality employed in such applications is not compelling, the larger normative project will lack sturdy foundations.

The third reason to be concerned with normative problems in a model of rationality is that unless we agree that the rational actors in behavioral models are making good decisions, the policy prescriptions that emerge from those models will not be persuasive. Overly stylized models of rationality tend to produce misleading welfare conclusions by neglecting to take important features of choice situations into account. For example, models of rationality that assume people are only concerned with income miss out on the important ways in which non-monetary factors also affect individuals’ well-being.³

¹ For Hobbes, citizens’ are bound to accept the legitimacy of government because doing so is the best way to satisfy their fundamental interest of self-preservation. Rawls’ arguments for the two principles of justice proceed from his claim that it would be rational for reasonable people in the original position to agree to them.
³ For example, unemployment may be harmful not just because of foregone income but also because of the feelings of worth that holding a job can bring. Similarly, education can be valuable for reasons other than increasing future earnings such as giving people the tools develop their preferences. See Sen (1999) for an application of these issues to human development.
Finally, an adequate normative conception of rationality is a necessary prerequisite for empirically testing whether people’s behavior is actually rational. One cannot conclude than an observed decision deviates from rationality without first establishing what it is that rationality requires. For example, to interpret experimental evidence of hyperbolic discounting as evidence of irrationality, it is first necessary to establish that hyperbolic discounting is normatively undesirable.

In this thesis, I aim to develop a model of rationality that is better equipped for these normative purposes. Beginning with a conception of individuals as having priorities they care about achieving, I understand rationality as a prescription for making decisions in ways that further those priorities. At a basic level, I take rationality to be concerned with the following question: “Based on my goals, what is it best for me to do?” More formally, a decision is rational when the decision-maker believes that the chosen option is at least as consistent with her preferences as any other available option. This is an instrumental conception of rationality in the sense that the rationality of a given decision depends solely upon whether the selected option furthers the priorities of the decision-maker.4

Beginning with this bare-boned understanding of what a good decision entails, Chapter 1 formalizes and develops this conception of rationality by addressing a range of foundational issues covering restrictions on the form and content of rational

4 It is important to distinguish my focus on the rationality of decision-making from the related issue concerning rational belief formation. Whereas the latter is focused on when it is rational to hold a particular belief, the former – my focus – is concerned with how people make decisions given their beliefs and expectations. Although a complete model of rational behavior requires addressing both of these issues, I limit my discussion to manage the size of the project. Developing a model for the rationality of beliefs would raise interesting questions about whether concerns other than accuracy should inform the belief creation process – questions whose answers are relevant for determining whether self-delusion or willful ignorance could ever be rational.
preferences, the relation between rationality and morality, and the role of higher order preferences. I also contrast my conception of rationality to the view informed by the revealed preferences approach to behavior, and argue that although that approach may be useful for gaining information about people’s concerns, it cannot be an effective foundation for normative analysis.

Chapters 2 and 3 pick up the model of rationality laid out in Chapter 1 and apply it to situations in which decision-makers experience self-conflict. In particular, Chapter 2 considers the dynamic conflict that occurs when a decision-maker expects to develop new goals that conflict with her current ones. Chapter 3 then turns to weakness of will, a form of self-conflict that decision-makers experience within a single point in time. Although cases of preference change and weakness of will have been approached descriptively with some success, they pose difficult questions for normative models.

After determining what rationality requires of decision-makers in situations characterized by these forms of self-conflict, Chapter 4 shows why decision-makers may find it rational to adopt strategies of self control, whether by precommitting to a particular course of action or by resolving to act in a particular way. I also demonstrate how strategies of self control can bolster one’s credibility in strategic interactions with others.

In my last chapter, I consider a variety of strategies that people may adopt to guide their decision-making other than the case by case optimizing method traditionally associated with rationality. Because it may be rational for people to make decisions in ways that result in non-optimal options being selected, such
strategies raise the possibility of decisions that are “rationally irrational.” After examining the viability of such decision-making strategies as tools of self control, I consider an application of these ideas to the relation between rationality and morality. Finally, my conclusion demonstrates the applicability of the concepts developed in this thesis by showing how they can help us approach some central questions in the social sciences.

From the get-go, it is important to distinguish between positive and normative models of decision-making. The understanding of rationality employed here is explicitly normative in that it is concerned with how people should make decisions, not how they do make decisions. In other words, a normative conception of rationality prescribes behavior rather than describing or predicting it. As such, when I refer to a particular choice as rational, I am not making a claim as to whether decision-makers will actually choose that way when faced with a decision, but rather that making that choice would be the best way for decision-makers to achieve their goals. Although the normative nature of our model deemphasizes the importance of how people behave, it also complicates our task because the questions raised here cannot be settled by reference to empirical evidence.\footnote{5 The distinction between positive and normative models of decision-making is especially important because of the large role that rationality plays in many positive economic models. These positive applications of rationality are formulated by first stipulating what behavior is rational and then by imposing the assumption that people make their decisions in rational ways. Although we will have occasion to critically assess the plausibility of this assumption – most notably in Chapter 3 – the task of formulating the underlying conception of rationality is conceptually prior; determining whether an option is rational does not require determining whether a decision-maker will choose it.}

The distinction between normative and positive can help us respond to a common objection leveled against broadly formulated models of decision-making, that what they gain in accuracy over simpler models, they sacrifice in usefulness. The
primary reason that theorists develop models of decision-making is to have a framework well suited for analyzing behavior. As George Box famously wrote, “…all models are wrong, but some are useful.” When a model grows too complex by seeking to incorporate too many factors, it defeats its broader purpose of providing a useful framework to analyze decision-making, or so the argument goes. This line of thinking implies that even if existing models of rationality are inaccurate, their inaccuracy is necessary if they are to remain useful.

The shortcomings of this argument can be illuminated by considering a less frequently cited insight also written by Box: “Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.” The problem with traditional models of rationality is not merely that they are wrong, but that their wrongness undermines their usefulness in important ways.

First, overly simplistic models of rationality yield prescriptions for behavior that are unpersuasive. It may be easier to determine the rational course of action by arbitrarily restricting the range of concerns that decision-makers might have, but the prescribed decision will not be compelling for people who have priorities other than the ones imposed by the model. For example, a decision-maker who cares about helping others would not be well served by making decisions in the way prescribed by a model that assumed she only cared about her own consumption.

Similarly, traditional models of rationality neglect to take into account features of choice situations that have relevance for which option is rational. For

---

example, a result that will emerge in Chapter 2 is that the rational way for decision-makers to deal with anticipated changes in their preferences depends upon why it is that they expect their preferences to change. A stylized model of rationality that prescribes behavior for such situations without considering this factor will not be useful because the results that emerge will be incorrect; choosing the option prescribed by such models would not be consistent with the decision-maker’s priorities. In general, a normative model of decision-making is only useful for prescribing behavior to the extent that decision-makers can expect the prescribed option to be the one that is in fact most consistent with their priorities.

Another pitfall with relying on overly simple models of rational choice is that their simplicity renders them inapplicable to important types of decisions that people are forced to face. As we shall see in Chapter 3, traditional models of rationality are unsuited for dealing with forms of self-conflict that decision-makers may experience, and are therefore unhelpful for analyzing behavior in those situations. As a result, they imply that it can never be rational for decision-makers to adopt strategies of self control to combat weakness of will – a claim we will challenge in Chapter 4.

Finally, it may turn out that the view of rationality we end up with is not the view that is easiest to model. Even if this concern motivates us to adopt a simplified model of rationality, it would still be useful to know what a less stylized model would consist of – if only so that we can distinguish the elements of the model that are imposed because they are normatively appealing from the elements that are imposed for modeling convenience. As such, arguments concerned with the tractability of the
model being developed do not indict the project at hand and can safely be placed on the backburner.

The question that we take on in this thesis – how to assess the quality of a decision – is a fundamental one, and the answers provided here will necessarily be incomplete. A model of rationality is not a self help book nor is it a practical guide for making decisions in our own lives. Nonetheless, the project of developing a normative model of decision-making will provide us with the tools and concepts that we need to approach such issues analytically. And along the way, we may even discover some insights that change the way we think about decisions in our own lives.
This chapter develops some of the fundamental tools and concepts we will use throughout the rest of the thesis. Section I defines and formalizes what it means for a decision to be rational. Section II examines two conditions that are frequently imposed on the form of rational preference orderings and questions whether preferences that violate these conditions are actually irrational. In Section III, I focus on the content of preferences and address the link between rationality and self-interest, before turning to the related topics of higher order preferences in Section IV and the relation between rationality and morality in Section V. Finally, Section VI contrasts the instrumental conception of rationality employed here to the revealed preferences approach pervasive in the economic literature, and argues that only an instrumental conception is adequately equipped for normative applications.

I. Basic Concepts

At a conceptual level, I define rationality to be concerned with decisions. By decisions, I mean situations in which people choose from a set of options that are available to them. These choices can be between physical actions (or inaction), as well as between more abstract states of affairs such as deciding whether to forgive a friend or whether to pay attention during a lecture.

Two features of this definition require clarification. First, the options must be mutually exclusive – only one may be chosen in a particular decision. For example, if I refer to a decision between a watermelon and a cantaloupe, I mean to imply that...
the decision-maker cannot choose both.\textsuperscript{8} Second, only conscious choices can constitute decisions. Many actions that we think of as voluntary are not the result of conscious choice, even when the actor would have been free to act differently had she considered the possibility of doing so. For example, slamming on the brakes to prevent a car crash or returning a slam in a fast-paced game of table tennis are not decisions because the action is not consciously selected, even if it could have been consciously controlled. Neither are habits such as turning off the lights when leaving a room, or other actions that people do unthinkingly like scratching an itch or biting a fingernail.\textsuperscript{9} In contrast to these cases, decisions refer to the conscious selection of options from a range of alternatives.

In any particular decision, the range of available choices is the decision-maker’s opportunity set. In order to assess the quality of these options, we need to be able to link them to the concerns that the decision-maker cares about achieving – her priorities. Most decision-makers have a variety of priorities, some of which may be more important than others. We can define the utility of an option as the degree to which that option is consistent with the decision-maker’s priorities.\textsuperscript{10} If a decision-maker believes one option to be more consistent with her priorities than another, the

\textsuperscript{8} Any set of options can be thought of in mutually exclusive terms by treating every possible combination as its own option. Thus if the decision-maker is not restricted in the number of fruits from which she may choose, we can say that her decision is between a watermelon, a cantaloupe, both, or neither.

\textsuperscript{9} Although any of these actions can emerge from a decision if the actor pauses to consider the action before taking it. If scratching an itch would be inconvenient, I can consider and choose from a range of alternative actions that might also fulfill my goal, such as thinking about something else while my hands are full.

\textsuperscript{10} More specifically, the crucial issue is the degree to which an option is expected to be consistent with the decision-maker’s priorities. Expectations are important here because a decision-maker cannot know with certainty the full set of consequences of the available options, and because such consequences are generally important determinants of an option’s consistency with one’s priorities. Only when a decision-maker’s concerns are exclusively deontological will expectations over alternatives not be relevant. As such, my use of the term utility is shorthand for expected utility.
The former has greater utility than the latter. For example, suppose that Evan’s only priority is that he eats as healthily as possible. If Evan were confronted with a decision between a steamed turnip and a grilled cheese sandwich, he would assign a higher utility to the steamed turnip than to the grilled cheese, assuming that he believed the steamed turnip to be healthier. In the context of Evan’s decision, we can say that steamed turnip is preferred to grilled cheese.

Before moving on, three points about utility and preferences deserve emphasis. First, utility is only meaningful in the context of a decision. At no point do people ever possess utility; it is a characteristic that can be associated with options, not with decision-makers. Second, both utility and preferences are only meaningful in the context of a particular decision. As we have defined them, preferences refer to rankings over available options, not to the states of the world associated with them. For example, saying that Evan has a preference for steamed turnip over grilled cheese is akin to saying that when presented with a choice between the two, Evan’s priorities would be better served by picking the steamed turnip. Finally, a third point of emphasis is that both utility and preferences are intrinsically agent-relative. The statement that one option is preferred to another is only intelligible in the context of a particular set of priorities and beliefs.

In many cases, two or more priorities may come into conflict; the option that would best satisfy priority A may not be the option that would best satisfy priority B. For example, Micah’s priority for healthiness may come into conflict with his priority...

---

11 Put differently, the preference relation is not defined over commodities, but over choices between commodities. This characterization is not crucial to most of our results, but will be helpful in our discussion of higher order preferences later in this chapter.
for tasty food. Were he confronted with a choice between a bland steamed turnip and a delicious grilled cheese sandwich, his decision would be harder than Evan’s. In cases of priority conflict, two factors determine which option has a greater utility. One consideration is the extent to which the options are consistent with the respective priorities. If Micah found grilled cheese to be only slightly tastier than steamed turnip, the relevance of his health priority would be smaller than if he loved the taste of grilled cheese and hated the taste of steamed turnip. The second factor relevant to resolving priority conflict is the relative magnitudes of the conflicting priorities. The more importance that Micah attaches to taste relative to healthiness, the more likely it is that he will prefer the grilled cheese to the steamed turnip. Finally, it is also possible for priorities to conflict in such a way that decision-makers are indifferent between two options. If the health benefits to the steamed turnip are exactly offset by the taste benefits to the grilled cheese, then neither option would be preferred to the other.

Taking advantage of the concepts developed thus far, we are now in a position to define rationality. We say that a decision is rational when the selected option is maximally consistent with the decision-maker’s priorities – that is, when the selected option is more consistent with one’s priorities than any other option that is also available.

For the sake of precision, we can represent this definition a bit more formally. Let \( A \) be the set of available options, and let \( \succ \) be a relation defined so that for two options \( x \) and \( y \), \( x \succ y \) if \( x \) is more consistent with the decision-maker’s priorities than \( y \). Additionally, we can define the maximal set, \( A^* \subseteq A \), that is made up of all the
maximal options: $A^* = \{a^*: \text{there exists no } a \in A \text{ such that } a \succ a^*\}$. In this context, we can characterize a decision as rational if the chosen option, $a^c \in A$, is a member of the maximal set, $a^c \in A^*$.

A feature of this definition that deserves emphasis is that the maximal set ($A^*$) may include more than a single option. If the decision-maker is indifferent between two options that are in turn preferred to all other available options, then both of those options are elements of the maximal set. Consequently, there may be more than one way to choose rationally when faced with a particular decision.

Another common way to define rationality is in terms of utility. Thus far we have been using utility to describe the degree of a preference for a particular option. As such, for a decision to be rational, its utility must be greater or equal to the utility of every other available option. A distinct concept is the idea of a utility function, which is a function that returns a numerical utility value for each element of the opportunity set. Options with greater utility values are preferred to options with lower ones. Because utility functions are derived from priorities, the option that maximizes the utility function will also be the option that is most consistent with the decision-maker’s objectives. Although utility functions are useful tools for analyzing decision-making because they make it easier to approach preferences mathematically, the ability to represent preferences through a utility function is not a precondition for assessing rationality. Although the conditions that allow a utility function may be
rationally necessary in their own right, the fact that certain preferences cannot be modeled with a utility function does not indict their validity in a normative sense.\(^\text{12}\)

As our definition makes clear, the concept of rationality is designed to focus on the relation between means and ends, rather than upon ends themselves. As David Hume famously put it, “‘Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger.”\(^\text{13}\) In practice, however, economists and other decision-theorists commonly impose restrictions on both the form and the content of preferences that can serve as the basis for rational choice. In the next few sections, I examine whether preferences that violate these conditions are actually irrational.

**II. Conditions on Rational Preferences**

Two restrictions on the form of preference orderings are almost universally imposed as requirements of rational preferences: transitivity and completeness.\(^\text{14}\) This section investigates the claim that incomplete or intransitive preferences are irrational.

How can these restrictions be justified in light of rationality’s explicit agnosticism regarding ends? I will consider two lines of attack as to why intransitive or incomplete preference orderings are irrational. The first argument is a practical problem with analyzing preferences that are incomplete or intransitive whereas the second argument questions whether such preferences can be interpreted in a

\(^{12}\) For example, lexicographical preferences cannot be represented with a utility function, but nothing about that class of preferences is obviously irrational otherwise.

\(^{13}\) David Hume, *Treatise of Human Nature* (1739), Book 2, Part 3, Section 3.

\(^{14}\) In addition to these two conditions, further restrictions are commonly imposed to cover choice under uncertainty, such as the independence axiom from the von Neumann-Morgenstern axiomatization of expected utility. In the interest of space, I will restrict my focus here to the more basic requirements of transitivity and completeness.
meaningful way. Before considering the justifiability of these requirements, I first review what it is that these conditions require.

For a decision-maker’s preferences to be complete, it must be the case that each available option can be ranked relative to every other available option. In other words, given any two available options \( x \) and \( y \), it must be the case that the decision-maker prefers \( x \) to \( y \), prefers \( y \) to \( x \), or is indifferent between \( x \) and \( y \). If none of these cases holds, then the two options cannot be ranked relative to one another and we say that the preference ordering is incomplete. Additionally, preferences are transitive if the following condition holds for any three available options \( x, y, \) and \( z \): if \( x \) is preferred to \( y \), and \( y \) is preferred to \( z \), then \( x \) must be preferred to \( z \).\(^{15}\) If this condition fails, then the corresponding preference ordering is said to be intransitive.

We are now in a position to consider the first argument as to why intransitive or incomplete preferences are irrational, namely that such preferences may be incompatible with rational decision-making. When preferences are intransitive, it may be the case that no maximal, or “best,” choice exists at all. If a decision-maker must choose between \( x, y, \) and \( z \), and \( x \succ y \succ z \succ x \), then every possible option is less preferred than another available option. Because the resulting maximal set is empty, every choice that the decision-maker could make would be irrational. In this way, intransitive preferences can preclude rational decision-making by preventing the emergence of a maximal option.

A similar argument can be made with regard to preferences that are incomplete. If two options cannot be ranked relative to one another, it may seem that

---

\(^{15}\) Mathematically: \( \forall x \forall y \forall z ((x \succeq y \& y \succeq z) \Rightarrow x \succeq z) \).
neither can be placed in the maximal set. However, the manner in which we have
been defining the maximal set mitigates the consequences of incompleteness. For an
option to be a member of the maximal set, the requirement is that no other option be
strictly preferred to it. To illustrate this point, suppose that both \( x \) and \( y \) are preferred
to all other available options, but that they are incomplete relative to one another.
This incompleteness does not disqualify them from the maximal set; that \( x \) is not
comparable to \( y \) is a quite different statement than \( x \) being less preferred than \( y \).
Consequently, incompleteness, unlike intransitivity, does not have the potential to
make rational choice impossible.

On the other hand, we could define rationality differently, with a stronger
standard for admittance into the maximal set. In particular, we could say that for a
decision to be rational, the selected option must be preferred to or indifferent with
every other available option. If we redefine rationality this way, there may be cases
in which incompleteness precludes rational decision-making in the same way that
intransitivity does. Because both of these definitions are plausible, we can adopt the
stronger one for the purpose of argument. If we then show that this stronger standard
of rationality can be satisfied when preferences are incomplete, then we will have
established the same result for the weaker standard as well.

We have seen that intransitivity and incompleteness can create situations in
which rational choice is impossible. For example, if a decision-maker’s preferences
are intransitive with respect to options \( x, y, \) and \( z \), and she is forced to choose from
among those options, then she would be unable to make a rational choice. However,
rational choice is only impossible when the intransitive options are preferred to every
other available option. If a fourth option \( w \) was also available that was preferred to \( x \), \( y \), and \( z \), then a maximal choice would be available, namely \( w \). In that case, the fact that the decision-maker’s preferences were not fully transitive would not preclude her from choosing rationally. In general, intransitive preferences only preclude rational choice to the extent that they prevent the existence of a maximal option.

The case is similar for completeness. Assuming that we use the strict standard of rationality defined above, we saw that incomplete preferences can preclude rational choice if the maximal option cannot be ranked relative to some other available option. Again, however, this difficulty emerges only to the extent that the preference ordering is incomplete \textit{relative to the maximal option}. When incompleteness only exists between non-maximal options, a rational choice can still emerge because the maximal option can be compared to every alternative. The logic behind this point is intuitive. It does not matter if I cannot decide whether I prefer getting shot in the foot or shot in the arm if I am also free to choose not getting shot at all. Despite the fact that my preferences are incomplete, a rational choice is still possible.

Given these qualifications, what are we left with? Incompleteness and intransitivity do not guarantee the existence of a rational choice, but they also do not guarantee that a rational choice will be impossible. It is tempting to impose completeness and transitivity on the preferences of rational decision-makers in order to guarantee the availability of a rational choice,\(^{16}\) the above discussion generates two misgivings about doing so.

\(^{16}\)It is especially tempting because unless preferences are complete and transitive, they cannot be represented by utility functions. Since the numerical levels of utility associated with each option \textit{are} complete and transitive – i.e., every real number can be ordered relative to every other number and for
First, these are stronger conditions than are necessary. Incompleteness and intransitivity only preclude rational choice to the extent that they preclude a maximal choice from emerging. In contrast, imposing only that a maximal choice is available in each situation would be necessary and sufficient to guarantee the possibility of a rational choice. Because an important goal of a theory of rationality is to accommodate the wide range of priorities that decision-makers may have, we should rule out preferences as irrational only when forced to do so.

Second, it is misleading to refer to incomplete or intransitive preferences as irrational on the grounds discussed thus far. In many cases, decision-makers whose preferences do not meet either of these conditions will be able to make perfectly rational decisions. Although completeness and transitivity guarantee the possibility of rational choice, their absence does not guarantee that such choice is impossible. The fact that a particular preference ordering may be associated with irrational decision-making is not reason enough to call the preferences themselves irrational. After all, any set of preferences has the potential to result in irrational behavior if the decision-maker who has those preferences happens to choose irrationally.

We have seen that the practical arguments in favor of the completeness and transitivity restrictions are insufficient. Although imposing these conditions is a convenient way to ensure that a rational choice exists, preferences that violate these conditions may still be compatible with rational decision-making and consequently,
should not themselves be termed irrational. In addition to the practical arguments in favor of the completeness and transitivity assumptions, another approach is to argue that there is something fundamentally irrational about preferences that violate these conditions.

Let us start with the completeness requirement. We are interested in whether it is fundamentally irrational for individuals to have incomplete preference orderings. To answer this question, we can distinguish between two possible reasons that a decision-maker might find herself with preferences that violate the completeness condition. First, incompleteness may reflect uncertainty on the part of the decision-maker over the content of her preferences. For many decisions, a good deal of effort must be spent not only on predicting the likely consequences of the available options, but also on determining the extent to which those consequences conform to one’s priorities, such as determining which of two effects is more important to a decision-maker. Until this effort is spent, the decision-maker will not know which of two options is more preferred. This behavior, in itself, does not seem irrational. The benefits to achieving a fully ranked set of preferences may be outweighed by the costs of putting in the effort required to do so. A point that will come up in Chapter 5 is that the cognitive effort required for such tasks may entail non-trivial costs. For example, a decision-maker may not know which of two options is more preferred, but still know that both are dominated by a third option. I may not know whether I prefer being shot in the foot or shot in the arm, but I have no reason to spend the effort ranking the two if I also have the option of eating some pizza. Similarly, a decision-maker may not know which of two options is her first choice, but if she expects the
utility of the two to be similar, it might not be worth it for her to invest the effort necessary to determine which is better. In these ways, a reluctance to expend cognitive effort explains how rational decision-makers can end up with preference orderings that are incomplete.

The above situation can be characterized as resolvable incompleteness because the preference ordering could be made complete if more effort were spent. The second reason that decision-makers might end up with an incomplete preference ordering is if some aspect of their preferences are themselves fundamentally incomparable. Such a situation might arise if agents believe that two of their options invoke incommensurate values. Although I want to avoid a discussion of whether incommensurable values can exist, if they do exist then it would follow that decision-makers who subscribe to them would have incomplete preference orderings. In that case, the source of the incompleteness would lie with the underlying values themselves, not with the decision-maker. As such, in this situation as well, it would be misleading to say that the incomplete preferences are themselves irrational.

We can now turn to the related argument about whether intransitive preferences are fundamentally irrational. The case against intransitive preferences being rational is more straightforward than the related argument concerning

---

17 Elizabeth Anderson illustrates the reasonableness of incommensurability between options with the following analogy: “Bach and Darwin were each highly successful in their own ways of being brilliant. Neither was superior in brilliance to the other, nor were they roughly equal in brilliance. If this were so, then a small but significant improvement in the brilliance of one would suffice to tilt the judgment in his favor. But it is silly to claim that, say, had Darwin achieved some brilliant insights into genetic theory as well as evolution, he would thereby have exceed Bach in brilliance” (56). In the same way, Anderson argues, options may satisfy decision-makers’ priorities in equally strong, but very different ways. Elizabeth Anderson, *Value in Ethics and Economics* (Cambridge, Mass: Harvard University Press, 1993), 55-64.
incomplete preferences. Suppose that a preference ordering is intransitive, so that \( x \) is preferred to \( y \), \( y \) is preferred to \( z \), but \( z \) is preferred to \( x \). Because preferences reflect the degree to which options are consistent with one’s priorities, this means that \( x \) is both more consistent with the decision-maker’s priorities than \( z \), but also less consistent with those priorities than \( z \). The idea that both of these statements can be simultaneously true is incoherent in a way that incomplete preferences are not. As such, it appears that intransitive preferences are fundamentally irrational.

As we noted at the beginning of this section, the claim that intransitive preferences are irrational is accepted almost universally, but it has occasionally been contested. I now consider one such objection raised by the philosopher Michael Philips (1989).

Philips defends the rationality of intransitive preferences by providing a variety of examples in which people make decisions according to intransitive preferences, but still appear to be acting rationally. By drawing on such situations, he concludes that intransitivity can arise rationally whenever decision-makers are motivated by multiple objectives, and must therefore not be irrational. I believe that Philips errs in both his explanation of the source of intransitivity as well as in his conclusion. I focus on his argument in some detail because its shortcomings illustrate important problems in the way that decision-making is often approached.

Philips offers several examples to demonstrate his point, but because they are structurally similar to each other, it will suffice for us to focus on only one:

**The Tennis Tournament**

Consider a tennis tournament that includes three players, A, B, and C. Suppose that my preferences for the outcomes ... are based on styles. Suppose that A is a power player, B is a traditional, graceful
baseline player and C depends on unorthodox, tricky shots. In that case I may prefer A’s winning the tournament to B’s winning the tournament because I prefer to see a power game defeat a baseline game (it is more daring); B’s winning the tournament to C’s winning the tournament because I prefer to see a baseline player defeat a tricky player (it is more aesthetic); and C to A because I prefer to see a tricky player defeat a power player (it is more comical). My preferences are perfectly rational but … they are not transitive.  

Philip’s example does seem to support his claim that intransitivities may arise rationally. A is preferred to B, B is preferred to C, and C is preferred to A. Yet nothing about the example seems implausible, and the preferences seem perfectly reasonable. Do we really wish to condemn such preferences as irrational just because they violate our transitivity condition? I will eventually argue that this line of reasoning is flawed because the preferences in the example are not actually intransitive. But that conclusion will be more apparent after some additional work.  

The first step on the path in responding to Philips is to note that transitive preferences can appear intransitive when they are conditioned on some changing external circumstance. To illustrate this, we can look at another example employed by Philips, drawn from a previous paper by von Wright:

**The Produce Party**

You are at a party and offered a choice between an orange and a banana. You choose the orange. You are then offered a choice between a banana and an apple and choose the banana. When offered a choice between an apple and an orange, however, you choose the apple.  

Are the partygoer’s preferences intransitive? An orange is chosen over a banana, a banana is chosen over an apple, but an apple is then chosen over an orange.

---


The reason this pattern of behavior does not seem irrational is that we can easily understand how it could have reasonably come about. For example, the decision-maker in the example could have chosen the apple over the orange because she had already received an orange that night and wanted to try a new fruit. The preference expressed in the third decision is different than the preferences expressed earlier because an important circumstance changed in between: receiving the first orange lessened the utility of the second available orange. When these pair-wise choices are considered independently, it appears that the decision-maker has an intransitive preference ordering. However, once the change in circumstances between the first and the third decisions is taken into account, it becomes clear that the preferences may not actually be intransitive. For preferences to be intransitive, it must be the case that the three pair-wise preferences all hold simultaneously. If the partygoer were suddenly confronted by a choice between all three types of fruit at once, intransitive preferences would be unable to produce a maximal option. Clearly, this need not be the case in the above example.

To emphasize the point even more conclusively, we can construct a hypothetical set of transitive preferences and show that they give rise to the same decisions made by von Wright’s partygoer. Suppose that our hypothetical guest prefers bananas to apples. He also loves oranges and prefers their taste to all other types of fruit, but has been informed by his doctor that his unfortunate vitamin C allergy limits him to one per night. Thus at the beginning of the party – before he has consumed any oranges – his preference ordering is *oranges > bananas > apples*. Thus when offered a choice between an orange and a banana, he picks the orange. But
after reaching his prescribed limit of vitamin C for the day, the guest strongly prefers to avoid oranges. As a result, his preferences become bananas > apples > oranges, and following this new ordering, he proceeds to choose the banana over the apple, and then the apple over the orange. The guest’s preferences are transitive both before and after the first orange is consumed, but because they change midway, they appear intransitive.

A similar problem arises in Philip’s tennis example. Here too, the apparent intransitivity of the preferences actually comes from the fact that the preferences expressed in each of these choices are contingent on changing circumstances. Unlike the produce party example – in which the changing circumstance had to do with the sequence of choices in time – what is changing in the tennis example is the set of options being considered. Each decision is framed in the context of a choice between only two players, and as these players change, so does the preference. In other words, the strength of the preference for any one player winning depends on which other player is being offered for comparison. For example, the preference for the tricky player is stronger when she is compared to the power player than when she is compared to the baseline player. If the circumstances in which the pair-wise comparisons are made is held constant, then the revealed preferences could be combined into an overall ranking. However, the preferences in this example are

---

20 Of course, a guest with preferences that really were intransitive would also make a series of pair-wise intransitive choices in this situation, but in that case, her choices would not seem as reasonable as those presented here. The irrationality becomes apparent when guests with intransitive orderings are forced to choose from all three types of fruit, and find themselves simultaneously believing the apple to be both more and less preferred than the banana.

21 This phenomenon – in which preferences depend on the other available options – is known as menu-dependence. Whether having preferences that change in this way is itself rational is another question, one which we will address in Section VI.
explicitly defined to change depending on which particular options are being considered; this feature is the source of the alleged intransitivity. As we saw in the produce party example, intransitivity only occurs when the three revealed preferences hold simultaneously. The decision-maker would be unable to answer the question, “Which of the three tennis players would you most prefer to win the tournament?” However, nothing about the example guarantees that this will be the case. Only when \( A \succ B, B \succ C, \) and \( C \succ A \) hold simultaneously will the preference ordering really be intransitive, and in that case, it will also irrational.

Philips anticipates a similar argument to the one I am making here. He points out that if the preferences expressed by every decision are interpreted as conditional on what other options were available, then the uniqueness of each pair-wise comparison would preclude our ability to synthesize pair-wise comparisons into overall preference rankings. However, the problem with synthesizing pair-wise comparisons into overall rankings only occurs when the decision-maker’s preferences are conditional on what alternatives are available. In other words, I am not arguing that changing circumstances always render pair-wise decisions incomparable, but only that when changes between circumstances are relevant – when they do affect one’s preferences – ignoring those changes will produce misleading results. Philips approach is flawed because he creates scenarios in which changing circumstances between decisions affect preferences, but then combines the preferences revealed by those decisions into an overall preference ordering which is assumed to hold over all circumstances.
Philip’s defense of intransitive preferences fails. It is true that the decision-makers in his examples seem to be behaving rationally, but this is because the preferences guiding their behavior are not actually intransitive. The prevailing intuition that intransitive preferences are fundamentally incoherent appears to be correct. However, the same cannot be said for the completeness condition. Not only does the practical argument against incomplete preferences fail, but people may have incomplete preference orderings for seemingly rational reasons. Thus we conclude that although completeness may be a convenient assumption to impose on rational decision-makers, it cannot be treated as rationally required. In other words, incomplete preference orderings are not fundamentally irrational in themselves, nor are they inconsistent with rational behavior.

In responding to Philip’s arguments, we have seen the importance of taking preference changes into account, as well as the dangers that changing circumstances pose for imputing preferences from observed behavior. This theme will emerge again throughout the reminder of this chapter, particularly in our discussion of revealed preferences in Section VI. But first, having completed our examination of the form that rational preferences can take, we can now turn to questions about their content.

III. Self-Interest and the Content of Preferences

An oft-cited and infamous characterization of rational economic actors is that they are excessively self-interested. To what extent is the model of rationality developed thus far vulnerable to the same criticism? And if our model of rationality is linked to self-interest, is this a weakness of the model? When one begins to investigate these
questions, it quickly becomes apparent that self-interest means different things to different people. In examining the relationship between rationality and self-interest, we can further our understanding about the content of rational preferences.

As we have defined the concept, decisions are rational when they are effective at advancing the priorities of the decision-maker. Whether rationality is tied to self-interest then, depends upon the content of the decision-maker’s underlying priorities. In particular, we can ask whether rational decision-makers may have priorities whose content is not linked to their self-interest. In answering this question, we can distinguish three possible interpretations of what priorities may be concerned with, each of which implies a different degree of support for the connection between rationality and self-interest.

The extreme self-interest view is the one traditionally associated with neoclassical economic theory. Two claims are crucial to this view; first, that rational actors make decisions to bring about their own happiness, and second, that each person’s happiness is unaffected by the happiness of others. The view of priorities implicit in this approach is that people’s concerns are entirely self-oriented – preferences for options do not depend upon the option’s effects on others. Because the standards employed in assessing rationality here are essentially self-oriented, this conception of preferences strongly ties rationality to the pursuit of self-interest. In

---

22 The concept of happiness is central here. Although I will not put forth a precise definition, I use the term to refer to mental states that the decision-maker considers desirable. More broadly, happiness is characterized by the presence of desirable emotions such as joy, relief, satisfaction, and the absence of negative emotions such as pain, guilt, dread, and so on. In this view, the claim that I would be happier in situation $x$ than in situation $y$ amounts to saying that I would prefer the mental state I would experience in $x$ to the mental state I would experience in $y$.

23 Except to the extent that the effects on others precipitate other effects that do affect the decision-maker.
this view, altruistic behavior – making decisions in order to promote the happiness of others – is fundamentally irrational, unless the agent expects to somehow benefit indirectly.

The extreme self-interest view of rationality described above has been usefully employed by generations of economists and plays a central role in many of the discipline’s most important results.24 However, the empirical assumptions that inform this view are suspect. In everyday life, we observe a large degree of behavior that suggests peoples’ happiness is directly affected by the happiness of others. Among many examples are parents’ concerns for their children, the willingness of people to help their friends, and the discomfort of television viewers when they see pictures of famine victims. These everyday observations have been confirmed by a large body of experimental evidence, in which participants consistently demonstrate a willingness to sacrifice personal gains for the benefit of others.

Some attempts have been made to explain away these apparent deviations from the extreme self-interest view by identifying hidden personal benefits that result from apparently altruistic behavior. For example, theorists have argued that self-interested decision-makers may choose to help others in order to make others willing to help them in the future. Although plausible for explaining some forms of apparent altruism, these explanations are not helpful for explaining altruism in situations where reputational effects are limited, such as one time laboratory experiments or anonymous donations. Others have argued that the assumptions behind the extreme self-interest view of rationality, although empirically suspect, are nonetheless

justified because they are easy to model. However, as discussed in the introduction, this type of consideration is of secondary importance for a normative model of rationality, in which it implies that decisions made on other grounds are normatively flawed. If theorists want to introduce these assumptions as simplifications, they should be explicit about doing so instead of treating them as axiomatic of rationality itself.

What these observations cast doubt on is the assumption that individuals’ happiness is not affected by the well-being of others. In response to these criticisms, some have conceptualized rationality more broadly to acknowledge the empirical fact that most people’s concerns are not exclusively self-concerned in the way assumed by the extreme self-interest view. This second view maintains the first claim of the extreme self-interest view – that rational decision-makers take actions to bring about their own happiness – but departs from the extreme view by relaxing the second assumption that their happiness does not depend upon the well-being of others. In loosening the restrictive assumptions that are often imposed on the content of rational preferences, this broad-happiness view of rationality can better accommodate people’s willingness to behave non-selfishly. For example, suppose that while walking down the street, I notice a twenty dollar bill fall out of the pocket of the pedestrian in front of me. I am faced with a choice between alerting the other person and keeping the money for myself. If I keep the money, I will spend it on a nice meal, from which I expect to derive pleasure. On the other hand, I may also expect to be plagued by guilt for not returning the money. If I decide that the pleasure from the meal would outweigh the negative feelings from the guilt, we could say that I would
be happier keeping the money than returning it. According to the broad-happiness view of rationality, the rational course of action would be for me to keep the money.

Although this model acknowledges that it may be rational to make choices in order to help others, the picture of rationality that emerges is still linked to self-interest in the sense that rational decision-makers only help others to the extent that they expect to benefit from doing so. In other words, it is only rational to help others when doing so will make one happier than not doing so. In some sense then, the rational altruism that emerges from this view is still selfish in a fundamental way.

The two models of rationality considered thus far differ in their assessment of what factors can affect one’s happiness, but they shared the basic focus upon happiness itself. In contrast, the goal-based view of rationality treats preferences more broadly. Instead of assuming that a decision-maker’s only goal is to achieve her own happiness, the goal-based view of rationality defines preferences to reflect consistency with any priority that a decision-maker may have. For most people, achieving happiness is an important priority, but it is usually not one’s only priority. For example, in addition to achieving happiness, decision-makers may also be concerned with acting morally. The distinction between the broad-happiness and goal-based views of rationality are related to the distinction that Amartya Sen draws between two forms of altruistic behavior: sympathy and commitment. Sympathy refers to people helping others because they expect to derive happiness from doing so, whereas commitment occurs when the motivational force of altruistic behavior is not predicated on achieving future happiness for oneself.
Because the goal-based view of rationality does not conceptualize priorities as exclusively concerned with one’s own happiness, it is less tied to self-interest than the other views we have discussed. Of course, it is still true that decision-makers in this view can only be rationally motivated based on their own priorities, but calling this self-interest would render the term applicable to any decision that a person makes on the basis of her own motivation. Consequently, the goal-based view of rationality is not intrinsically linked to self-interest unless one defines self-interest to include any purposeful behavior, a definition that would severely erode the meaning of the term.

The goal-based view of rationality allows for people to be rationally motivated by concerns other than achieving happiness. Whether we accept this view depends upon whether we believe it can be rational for people to have priorities other than the pursuit of happiness. If we do not, then we are forced to conclude that it is irrational for a decision-maker to select an option when she believes that a different option could have achieved more happiness for herself. Because we want to develop a broad conception of rationality that can accommodate the fact that reasonable people have a broad range of concerns, our presumption is adopt the goal-based view. However, before doing so, it is necessary to consider several potential objections to the broader approach.

First, one might argue that priorities other than seeking happiness are themselves fundamentally irrational. Choosing to restrict the realm of rational concerns to anticipated mental effects seems arbitrary and I know of no good argument to back it up. Another concern is that expanding rationality in this way would make it unsuitable for modeling purposes, but this approach is no more
difficult to model than the happiness-seeking view described above. After all, anything that might be relevant to one’s priority may also be relevant to one’s mental-state. Furthermore, as I argued above, tractability considerations – although important – are of at most secondary importance when formulating a normative model. What is at issue between the second and third approaches is not so much a difference in whether any particular behavior could be rational, but is more about how we interpret what it means when someone acts rationally, within a broader normative framework.

The third objection to the goal-based version of rationality is the one I encountered most when describing this idea to others. In general, the argument runs something to the effect that people are only motivated to do something when they expect that doing it will make them happier. This claim has some plausibility because priorities and mental-state considerations often overlap. For example, many people who have the priority to help others also derive pleasure from altruistic actions. However, this objection can only be persuasive if it succeeds in demonstrating that priorities can have no motivational force outside of happiness considerations. I believe that it fails to do so, and employ two thought experiments to support my case.

The first thought experiment draws on an idea proposed by Jon Elster. Consider why people choose to act morally when doing so constrains their choices in inconvenient ways. An argument along the happiness-seeking version of rationality is that they do so in order to avoid the guilt that they would feel if they chose to take

immoral actions. However, suppose that a new pill were invented, which when consumed, would erase the guilt felt from a prior action. Would people choose to take these pills to free themselves from the constraints of morality? For decision-makers who choose not to, there must be some motivational force apart from mental-state considerations pushing them away from the immoral action.

A more developed example can be found in the work of Robert Nozick. Consider his experience machine, a device designed to provide people with any experience they could possibly desire, for instance writing a great novel, making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. While in the tank you won’t know that you’re there; you’ll think it’s all actually happening. Would you plug in?

This thought experiment nicely gets at the inadequacy of a mental-state approach to priorities. Choosing to plug in would certainly bring about the best mental states – it would guarantee happiness and one would have no idea that the pleasurable experiences were only simulated. Consequently, choosing to plug in would be the rational choice according to the happiness-seeking view of rationality. On the other hand, it is easy to imagine that some people would be deeply reluctant to plug into the machine, expected happiness notwithstanding. Perhaps these people care about accomplishing some goal in the world, or living an authentic existence, or have some other priority along these lines. Unless we are willing to label anyone who decides against plugging in as irrational, we must also acknowledge that people are capable of being rationally motivated by concerns other than the pursuit of desirable

---

mental states. Because the happiness-seeking model cannot take such considerations into account, it is deeply inadequate as a conception of rationality.

Another problem with the objection to the goal-based view is that the assumption that people have no concerns other than the pursuit of happiness is essentially an empirical claim about the content of people’s priorities. Our two thought experiments cast doubt on the accuracy of this claim, but even if it were true as a descriptive fact, it should not be taken as the basis for a normative model of rationality. In other words, the question we are interested in when designing a normative model is not how people do behave, but how people should behave. If it turns out that the decision-makers to whom the model is applied do not have priorities other than the pursuit of happiness, then the goal-based view reduces to the happiness-seeking approach. As such, unless we are persuaded that holding other priorities would be intrinsically irrational, we have no reason to exclude such motivations from the framework of our model a priori.

In this section, we have seen that an adequate conception of rationality – capable of accommodating the broad range of concerns that reasonable decision-makers may hold – requires a movement away from the narrow happiness-seeking and even narrower extreme self-interest approaches that have traditionally been associated with rationality. People’s concerns and priorities are not limited to furthering their own happiness. And for decision-makers with broader priorities, the rational course of action may often be to act against one’s own self-interest.

IV. Higher Order Preferences
We have seen that a person’s preferences may encompass a multitude of concerns, including goals that stem from priorities other than the self-interested pursuit of happiness. One interesting category of preferences not yet discussed is second-order preferences, that is, preferences about one’s own preferences. For example, a smoking addict may have a preference for smoking because of the pleasure that cigarettes provide, but simultaneously prefer that he did not have that preference because he recognizes that smoking is unhealthy. The former is a first-order preference and the latter is a second-order one. In addition, people may have third-order preferences about their second-order preferences, and so on. More generally, we can refer to all preferences whose content concerns other preferences as higher order preferences, and contrast those to first-order preferences whose content does not involve other preferences.

The claim that people have higher-order preferences is not particularly controversial in itself. Economists ranging from Amartya Sen (1977) – who prefers the term “meta-preferences” – to John Kenneth Galbraith (1958) have recognized that decision-makers can critically reflect on their preferences and attempt to shape them as they see fit. The philosopher Harry Frankfurt (1971) has argued that the capacity for critical reflection about one’s desires is at the very core of personhood, and is what separates persons from other actors such as animals and what he calls unreflective wantons. A question that has been less extensively addressed is the degree to which instrumental models of rationality can accommodate second-order preferences, or whether this type of concern demands an entirely different
framework. In particular, it is not apparent how an instrumental conception of rationality focused around achieving one’s goals can accommodate preferences whose goal is to change the decision-maker’s goals themselves. In this section, I explore this issue by attempting to integrate second-order preferences into the model of rationality developed thus far.

In approaching this issue, we can draw a useful distinction between two terms that until now have been mostly conflated: priorities and preferences. Priorities refer to decision-makers’ broadest concerns about the world, objectives that are valued for their own sake. In contrast, preferences refer to rankings over individual options. Preferences are derived from priorities; they reflect the degree to which a particular option is consistent with the decision-maker’s priorities. The strength of a preference for an option depends upon the degree of consistency between the option and the priorities, as well as the importance of the priorities that are invoked.

An example can help clarify the relation between priorities and preferences. Suppose that I am hungry and must decide between eating chocolate cake (C) and Brussels sprouts (S). As someone who loves chocolate and dislikes the taste of Brussels sprouts, I believe that choosing C would be more consistent with my priority of achieving happiness – it would feel better. On the other hand, I also know that Brussels sprouts are much healthier for me than chocolate cake, and so S is more consistent with another of my concerns, my priority for longevity. In choosing between S and C, a number of considerations are relevant. The strength of my preference for S is a function of S’s consistency with my happiness priority, its

\[27\text{ Although see George (2001) for an exception.}\]
\[28\text{ Or rather, the decision-maker’s belief about this degree of consistency.}\]
consistency with my longevity priority, and the relative importance of the two priorities. Similarly, my preference for C is a function of C’s consistency with the two priorities and the invoked priorities’ relative importance. For example, if I attach substantially more importance to happiness than to longevity, then it is likely that my preference for C will be greater than my preference for S. Alternatively, if I discover that the health-risks from eating the cake are even larger than I expected, then my preference for C will decrease and it is more likely that I will prefer S to C.29

Having distinguished between preferences and priorities, we are now in a good position to take on second-order preferences. Like any other type of preference, second-order preferences are derived from a decision-maker’s priorities. They arise in situations like the above example in which the available options invoke conflicting priorities. In the example, my priority for happiness clashed with my priority for longevity. Suppose that all considered, I prefer C to S. This means that in situations where I am confronted with the two options, it will be rational for me to choose C over S. Because this preference has negative consequences on my health, my priority for longevity would be better served if my preference ranking was reversed – that is, if I preferred S to C. If this were the case, then I would not choose to eat the cake when given the opportunity and would thus avoid the damaging health effects from doing so. As a result, we can say that my priority for longevity gives rise to a second-order preference that I prefer S to C. Although my first-order preference is for C over

29 I must confess that my attitude towards Brussels sprouts have changed since writing this example. It turns out that they are actually quite tasty when roasted with oil, salt, pepper, and lime.
S (C ≻ S), I also have a second-order preference that I prefer S to C:
((S ≻ C) ≻ (C ≻ S)).

Having seen how second-order preferences can arise, we may ask what rationality requires in those situations in which first- and second-order preferences diverge. I believe that the rational course of action in such situations is to follow the first-order preference. After all, rationality is fundamentally concerned with furthering one’s priorities in the world, and the fact that one has a first-order preference for a particular option implies that that option is most consistent with one’s priorities. A conflicting second-order preference does not negate this fact; it only indicates a decision-maker’s desire that this were not the case, and that some other option were in fact more consistent with her priorities. As such, the mere presence of an unfulfilled second-order preference does not change the rationality of acting according to first-order preferences.

If second order preferences do not directly conflict with first-order preferences, where do they fit into a model of rationality? Second-order preferences come into play when decision-makers have the option of altering their first-order preferences. To see this, recall that a preference for an option reflects the degree to which that option is consistent with the decision-maker’s priorities. As such, a change in preferences can be brought about by altering the degree to which options are consistent with priorities. To illustrate, return to our chocolate cake example.

Picking C best satisfies my happiness priority because eating cake makes me happier.

---

At this point, one could plausibly object that we have only presented half the story, and that my happiness priority might give rise to a conflicting second-order priority in the opposite direction. This is not the case, but it will take a little more work to see why.
than eating sprouts. If chocolate cake ceased tasting good to me, eating it would no longer be consistent with my priority for happiness, and I would prefer choosing the healthier option. Thus my second-preference for preferring Brussels sprouts to chocolate cake gives me reason to try to modify the degree to which the options are consistent with my priorities. In particular, it motivates me to try to decrease the happiness I feel from eating chocolate cake and to increase the happiness that I feel from eating Brussels sprouts. If I can succeed in making S more consistent with my priority for happiness than C, I will no longer prefer C to S and my second-order preference will be fulfilled. In other words, altering the options’ consistency with my priorities eliminates the conflict between the two priorities, both of which now point to choosing S.

We are now in a position to see why my happiness priority does not give rise to a conflicting second-order preference that I prefer C to S. My happiness priority only prescribes selecting C over S to the extent that I enjoy the taste of cake more than the taste of sprouts. If this state of affairs is reversed, and I come to believe that eating sprouts will make me happier than eating cake, then my happiness priority would prescribe selecting S instead. In other words, the priority only supports picking options that do make me happy; it is neutral with regard to which options will best accomplish that task.³¹
To see how first- and second-order preferences can interact, suppose that I happen across a hypnotist who conveniently specializes in reversing food preferences. She promises that in exchange for $30, she can ensure that I stop enjoying the taste of chocolate cake and begin to crave the taste of Brussels sprouts. To decide whether to undertake this service, I need to compare the strength of my second-order preference to my first-order preference for keeping the money. Determining which of these preferences is stronger is essential for deciding whether it is rational to accept her offer. Although hypnosis may not always be a viable approach to bringing about changes in preferences, the issues raised by this example are important for understanding participation in markets for preference-changing technology, such as wine appreciation classes or drug treatment programs.

Luckily, second-order preference changes can often be achieved without relying on intervention from third parties. People reflect on their tastes and cultivate them in a variety of ways. Repeated listening to classical music is an oft-cited technique for coming to enjoy it more, just as getting used to an unpleasant task can make doing it more bearable. For addictive behaviors, repeating the behavior strengthens the addiction and increases future repetitions’ effects on happiness. Consequently, undertaking or refraining from an addictive behavior can often be a tool to achieve preference change. For example, smoking addicts that have a second-order preference against wanting to smoke can achieve their second-order preference by refraining from smoking and breaking their addiction. Similarly, someone considering whether to try heroin for the first time may be dissuaded out of a second-order preference against developing a preference for heroin. In these ways, second-
order preferences can be relevant to any decision in which one of the options affects one’s future preferences.

Although second-order preferences are usually discussed in the context of changing tastes, such as one’s appetite for a given food or desire for a cigarette, not all second-order preferences are self-oriented in this way. For types of priorities not oriented around the decision-maker’s own mental response, preferences can be altered by changing circumstances external to the decision-maker. In other words, there are cases in which the basis for preferences – the degree of consistency between options and priorities – depends upon circumstances other than the decision-maker’s tastes. For example, suppose that I enjoy a particular brand of coffee, but find myself morally opposed to the working conditions in which the coffee is produced. After taking my conflicting priorities for happiness and morality into account, I judge that I prefer not buying the coffee (N) to buying the coffee (B), \( N \succ B \). However, in the same way that one’s priority for longevity can produce a second-order preference against chocolate cake, here my priority for happiness produces a second-order preference in favor of buying the coffee. In particular, if I can improve the conditions in which the coffee is produced, I can alleviate the tension between the two priorities and reverse my preference so that I can enjoy the coffee without believing it to be immoral. In the same way that my second-order preference against the chocolate cake gave me an incentive to undergo hypnosis, my second-order preference for buying the coffee provides me with an additional incentive to work to change the conditions in which the coffee is produced. In this way, decision-makers’ preferences for states of the world may result directly from priorities about those conditions as
well as indirectly through second-order preferences, from other priorities that could be better fulfilled if conditions were different.

Before concluding this section, I want to discuss another strategy through which second-order preferences might be realized. Because preferences reflect the consistency of options with priorities, we saw that second-order preferences may be satisfied by changing the degree to which options are consistent with priorities. However, this approach is often ineffective because decision-makers have imperfect control over the circumstances that determine priority-consistency; most people cannot change the fact that they enjoy the taste of chocolate cake and the average coffee drinker lacks control over the conditions in which coffee is produced. These shortcomings suggest that an alternate approach to satisfying second-order preferences would be more effective.

An alternate strategy to accomplish preference change is to modify one’s underlying priorities themselves. For example, rather than bringing about a preference for coffee by changing the degree to which drinking coffee is consistent with my priority for moral action, I can also bring about the new preference by increasing the relative importance of my priority for happiness relative to my priority for morality. The priority change approach to second-order preferences is appealing because there is a sense in which one’s priorities are entirely under one’s own control. If I do not want to attach importance to a certain issue when making my decisions, then that issue ceases to be one of my priorities. Similarly, I may be able to make a

---

32 Additionally, there may also be situations in which adopting an entirely new set of priorities may help satisfy my existing priorities. We will investigate such cases in Chapter 5.
new issue into a priority simply by deciding that I want to satisfy it when making my decisions.\footnote{Changing one’s priorities to satisfy a second-order preference is different from another way that priorities might change, which is from a change in judgment. For example, after reading Kant, I may decide that I should make not deceiving people into a priority or I might make not hurting animals into a priority after reading Peter Singer. These changes are different because they stem from a change in judgment, rather than a conscious decision to adopt a new priority for the sake of one’s current priorities.}

Although this approach may be possible when other strategies for changing preferences are not, it suffers from an important problem. The defining characteristic of a priority is that its satisfaction is intrinsically valuable; people care about satisfying priorities as an end in itself. When a new priority is adopted in order to satisfy a second-order preference, however, its value can only be instrumental. In other words, it can only motivate decision-makers to the extent that decision-makers see its satisfaction as valuable, that is, as long as it satisfies their original priorities. Why is this so bad? It means that the value one attaches to it – its status as a priority – is contingent on it being consistent with the other, original priorities. Thus decision-makers only care about fulfilling the new priority to the extent that its fulfillment is consistent with their old priorities. Because the new priority offers no unique motivational force, this strategy cannot be an effective method for satisfying second-order preferences.

One way around this dilemma would be if the very act of adopting a new priority could provide it with intrinsic value. The issue here is how people assign value. Although an adequate treatment of this question would extend far beyond the scope of this thesis, we can sketch an answer here. For a priority to have motivational force, it must seem like an important reason for acting when the decision-maker
reflects upon it. For most people, the source of this importance comes from believing the priority to fit into a broader narrative about the world, such as religion, culture, or a secular philosophical system. For example, people who have priorities against lying – that is, people who see lying as intrinsically wrong and something to be avoided – typically derive this value from a religious imperative, inherited cultural norms, or a philosophical system of ethics. For people whose priorities are derived in such ways, simply deciding to change priorities will not be enough. A devout Christian, who believes that the Bible prohibits lying and that the Bible is the best source of value, cannot simply decide to adopt a priority in favor of lying because the new priority will seem unpersuasive whenever it is reflected upon. In general, priority change is ineffective to the extent that people’s priorities are grounded in a broader system of value in which the source of value depends on more than what is convenient for the decision-maker in question. For most people, a system of value would be overly arbitrary if the mere act of choosing to adopt a particular goal is enough to make the pursuit of that goal valuable.\textsuperscript{34}

This section incorporated second-order preferences into our model of rationality by showing how second-order preferences are fundamentally about states of the world. Preferences reflect the consistency between options and priorities, and that consistency is simply a state of the world that people may have preferences about. The factors that determine consistency between options and preferences can

\textsuperscript{34} This criticism is not universal because it only holds for decision-maker whose priorities come from systems of value that do not attach importance to a goal simply because the individual chooses to pursue that objective. Existentialist writers such as Sartre have proposed systems of value in which individual choices to adopt a particular pursuit is the only compelling source of value. For decision-makers who adopt such systems of value, priority change may be an effective tool to satisfying second-order preferences. To adequately address such issues, a fuller discussion of existentialism would be necessary.
either be internal, such as one’s taste for Brussels sprouts, or external, such as the working conditions in which coffee is produced. By changing these circumstances, decision-makers can expand the number of priorities that can be satisfied simultaneously. When second-order preferences are conceptualized in this way, incorporating them into models of rationality is straightforward. The benefits of satisfying a particular second-order preference can be weighed against the cost of bringing about the state of affairs necessary to achieve the desired change in preference.

By demonstrating that our conception of rationality can accommodate second-order preferences, we have shown that instrumental models of rationality can be applied to an important class of concerns that have not been traditionally considered. Decision-makers’ preferences do not need to be taken as given and attempts to change preferences can be analyzed within the same framework as other types of decisions. At the same time, our discussion of the shortcomings of intentional priority change hinted at where a limit to rational analysis may lie. Although decision-makers may have rational reasons to adopt new priorities, in general, the new concerns cannot be intrinsically valuable; they cannot be priorities themselves. This result suggests an understanding of priorities as the starting points from which practical rationality proceeds. Such a characterization does not deny that it is important to assess the quality of people’s underlying concerns, only that attempts to do so are cannot be grounded on appeals to rationality. In the next section, we consider another basis for prescribing behavior that may be better suited for this task.
V. Rationality and Morality

We have seen that an adequately broad conception of rationality leaves room for decision-makers to have priorities of a moral nature. Just as people can be concerned with achieving happiness, so too can they be concerned with acting morally or bringing about just states of affairs. To the extent that people’s priorities are of this type, moral considerations can play an important role in rational decision-making.35

At the same time, the capacity of rationality to reflect moral concerns does not negate the fact that rationality and morality differ in important ways. Both morality and rationality are ways of prescribing behavior, in that they both answer the question, “What should I do?” However, the manner of determining the prescription differs between the two. The answer to what choice is rational depends on the priorities of the decision-maker in question, whereas, we generally think of the moral course of action as being determined by external, ethical considerations.

Although the process for coming up with the behavioral prescription differs between rationality and morality, there may well be instances in which both prescribe the same course of action. For example, people typically feel guilty after doing something immoral just as they may experience satisfaction from doing what they believe to be right. Because of this overlap, it will be rational for decision-makers to act morally out of concern for their own happiness. On the other hand, although

35 One might object that some types of moral concerns, specifically deontological constraints on action, cannot be adequately represented as a priority like any other subject to maximization. It could be argued that this approach misrepresents deontological moral concerns, which should be seen as absolute constraints on action. This problem can be fixed by representing the deontological concerns as priorities with infinite weights, which effectively transforms the resulting preference into a constraint on action; courses of action that violate these constraints are always less preferred to morally-acceptable courses of action. The formal equivalence of these two approaches is demonstrated in Sen, *Rationality and Freedom*, 196-197.
emotional forces have the capacity to bring rationality and morality into alignment, there may still be situations in which the two provide divergent prescriptions.

To illustrate how rationality and morality may diverge, consider a scenario proposed by Kurt Baier, in which one is offered a choice between two Mayan vases. “One is very beautiful, in very good condition, relatively inexpensive, and almost certainly supplied by professional grave robbers, while the other is less perfect, has suffered some damage, is larger and considerably more expensive, and has almost certainly been excavated and imported legally.” 36 This situation can be analyzed from either a rational or moral perspective, and the prescription may well differ between approaches. Considerations that may be morally relevant, such as whether one believes grave robbing to be theft, may be less important from the perspective of rationality, which may also emphasize factors such as the difference in quality and how much guilt the decision-maker will expect to suffer. In general, it seems that the coincidence of rationality and morality will depend upon the specific circumstances of the decision under consideration.

What emerges from this conclusion is a fundamental disconnect between rationality and morality. There may be cases in which people “should” act immorally – when “should” is given weight by rationality – and likewise, there will also be cases in which people “should” act irrationally – this time when the “should” statement is given force by morality. To illustrate the former, we can think of cases in which people do not attach much importance to acting morally. In turn, the latter case can be illustrated by people who have priorities that we do not think of as moral, such as

white supremacists, for whom the rational course of action – the one which best furthers their objectives – may be radically different from the option we believe to be most moral. As such, to say that it is not rationally required for some people to be moral because their priorities do not dictate moral behavior is not to say that they should not be moral, or that their immorality should be sanctioned. Describing a behavior as rational only endorses it to the extent that the decision-maker’s priorities are endorsable on their own, and as we have seen, that task is outside the bounds of rationality.

In addition to the theoretical possibility that morality and rationality can produce divergent prescriptions, a deeper source of difference between the two prescriptions can be found in the source of their authority. A central theme of moral theory is that moral requirements on behavior should be inescapable – that is, the imperatives on behavior produced by morality should be categorically binding. The importance of this theme is central to Kant’s moral theory, in which he argues that the motivational force of morality is categorically binding because people must conceptualize themselves as acting freely. In contrast, Hobbes’s moral theory is often criticized on the grounds that its motivational force stems from a desire he supposed to be universal and primary – the desire for self-preservation. As a result, the motivational force behind Hobbes’ ethical system is contingent upon people prioritizing the objective of self-preservation. As Nagel puts it, “[Hobbes’] ethical system is just the development of what is rational for people who have that end … the basic motive is taken as given, and only its consequences qualify as ethical
considerations.” For people who do not share the exclusive focus on self-preservation, Hobbes’ ethical prescriptions are escapable.

An argument against rationality serving as the basis for morality can be advanced along these lines. Just as the authority behind Hobbes’ prescriptions stems from the supposed universality of the priority of self-preservation, rationality can only prescribe a particular behavior based on the priorities of the decision-maker. What is rational for one decision-maker will often be irrational for a different decision-maker with different concerns. The fact that rationality is contingent upon a decision-maker’s priorities means that it can only be a hypothetical imperative, not a categorical one.

The criticism I have just outlined is certainly far from universally accepted, and many since Kant have attempted to argue that a thicker connection between rationality and morality does exist. Although my purpose in this section is only to outline the conceptual relation between rationality and morality, I will sketch a few of the approaches one could take to arguing that the connection is stronger than I have painted it here. One approach would be to restrict which ends are rationally acceptable, and argue that that only the pursuit of moral ends can be rational. For example, Nagel argues that only certain types of agent-neutral reasons can adequately serve as acceptable ends and Macintosh (1998) claims that only moral ends can serve as the basis for a rational set of preferences. Although I will not go into the specifics of these theories, I believe that the standards they employ for rational priorities are excessively strict.

The other approach to linking rationality and morality is to accept that individuals may happen to hold non-moral ends, but to claim that the only rational way to pursue those ends is through moral action. The role of Hell in Christianity can serve this function because the consequences of immoral action are so bad that they overwhelm any other motivations that potential sinners might have. A more subtle argument along these lines is offered by David Gauthier, but we are far from having the tools necessary to investigate it. We will return to Gauthier’s argument in Chapter 5.

VI. Revealed Preferences

The foundations for our theory of rationality are nearly in place. We have examined a variety of possible ways to conceptualize the ideas at the heart of the model, and have touched upon some of the important differences between rationality and morality. Before we take up the task of applying our model to issues of self control – the project that will occupy the rest of the thesis – the final section of this chapter emphasizes some of our model’s defining features by contrasting it to a quite different conception of rationality: the revealed preferences framework. After describing some of the differences between the two approaches, I argue that models of rationality founded upon revealed preferences are unsuited for normative analysis.

Theories of rationality founded upon a revealed preferences approach are quite different from our model because they start from a radically different understanding of preferences. Rather than conceptualizing preferences in terms of a person’s priorities and beliefs – as we have done – the revealed preferences approach
takes preferences to be descriptions of behavior. In that view, saying that a decision-maker has a preference for star fruit over kiwi means that she will choose star fruit when confronted with a choice between the two.

This account of preferences is simplifying in two ways. First, on an epistemological level, it simplifies the task of determining the content of people’s preferences. Because preferences are information about how a decision-maker will choose in a particular decision, the choice that is made in that decision by definition reveals the content of the preference. Second, the revealed preferences approach simplifies the task of determining whether a decision is rational. Because decisions reveal preferences, if the decision-maker chooses differently in two repetitions of the same choice situation, she must have contradictory preferences. In at least one of those instances, the decision-maker must not have chosen most consistently with her preferences. This observation simplifies the empirical task of determining whether a decision-maker is acting irrationally because assessing rationality only requires that current decisions be evaluated in light of past decisions, rather than invoking the decision-maker’s goals and beliefs. It is in this sense that proponents of the revealed preferences approach claim that rationality involves only internal consistency of choice, rather than consistency with factors that are external to demonstrated behavior, such as one’s priorities.

Given these benefits, it may appear that a revealed preferences conception of rationality is superior to the instrumental approach that we have been focusing on thus far. In this section, I challenge that claim first by showing that internal

---

38 Assuming, of course, that her preferences did not change.
consistency is not enough to assess rationality, even within the revealed preferences framework, and second, that the revealed preferences approach is ill-suited for normative applications.

The standard of internal consistency is more complicated than it first appears. In particular, assessing internal consistency requires imposing some external judgment about when two decisions are equivalent. A few examples shed light on this issue. Consider a variant of the Produce Party example from Section II, in which a guest is first offered a choice between a mango and a peach, and then, after choosing and eating the mango, decides on a second occasion to pick the peach. If the only relevant features of the situation are the available options, then the two decisions would be equivalent. Consequently, it would appear that the guest is choosing inconsistently and is therefore irrational. However, this interpretation seems silly because a relevant feature of the situation has changed: the partygoer has already consumed one mango.

At this point, one could plausibly ask how we can decide whether a particular change in circumstances is relevant. What matters for determining relevance is whether the partygoer’s preferences are contingent on the changing feature of the situation. If the partygoer’s preferences between the two fruits are dependent on what she has eaten in the recent past, then the two choice-situations are not equivalent. We are thus unable to decide whether the two decisions are equivalent without knowledge of whether the partygoer’s preferences are contingent on the changing circumstance. Thus we see that even in a revealed preference conception of rationality, identifying irrationality still requires knowledge about the decision-maker’s goals.
One might respond that allowing such changes to determine equivalence makes rationality impossible to evaluate. Because decision-makers may have preferences that are contingent on anything, and because at least one feature – i.e., time – changes between most observable decisions, it is never possible to conclude definitively that an observed inconsistency reflects irrationality. Although, I accept that this observation is accurate, it may still be possible to conclude that some observed inconsistencies probably reflect irrationality. Consider a famous experiment conducted by Daniel Kahneman and Amos Tversky:

**The Asian Disease Experiment**
Groups of respondents were asked one of the following questions:

**Question One**
Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed.
If Program A is adopted, 200 people will be saved.
If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

**Question Two**
Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed.
If Program C is adopted 400 people will die.
If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

In Question One, a strong majority of respondents (72%) chose Program A. In Question Two, an even larger majority (78%) chose Program D. What is notable about these results is that apart from the wording, the two questions are exactly identical! Program A and Program C have the same consequences, but one is framed

---

negatively and one framed positively. The same is true for Programs B and D. Yet respondents preferred the less-risky program in Question One and the riskier program in Question Two.

Does this inconsistency imply irrationality on the part of the respondents? After all, the two situations seem to elicit contradictory preferences. As we noted above, however, we can call two decisions equivalent only if the preferences of respondents were not contingent on the difference between the two situations. And although it seems unreasonable, it is not impossible that respondents had priorities that were contingent on the wording of the questions, which is to say that their priorities were better served by avoiding risk when the choice was worded one way and by embracing risk when the choice was worded differently. Consequently, we can only interpret this experiment as evidence of irrationality if we are willing to impose the additional assumption that participants did not have preferences that were contingent in this way. In this case, we can be confident in making the assumption because the assumed-away preference appears particularly implausible. However, the necessity of imposing this assumption about the content of the participants’ preferences illustrates how internal consistency is insufficient, even in extreme cases such as this. In other words, we can only be confident in our judgment of irrationality to the extent that we are confident about the veracity of our assumptions about the decision-maker’s goals.

This conclusion extends to all attempts to identify irrational behavior through the revealed preferences approach. Interpreting observed inconsistencies as evidence of irrationality necessitates a willingness to impose assumptions about the content of
people’s preferences, namely that they are not contingent on features that change between the two observed decisions. Thus we see that even in the revealed preferences framework, we must still fall back on external factors in order to assess rationality.

These insights can help us examine the validity of a central assumption of the internal consistency conception of rationality, the weak axiom of revealed preferences (WARP). First put forward by Samuelson (1938), WARP stipulates that if a rational decision-maker chooses \( x \) when both \( x \) and \( y \) are available to her, she will not choose \( y \) on another occasion when \( x \) is also available. In other words, if a choice reveals one option to be preferred to another option, then subsequent choices in which the same two options are available will not reveal the opposite preference, at least when people are choosing rationally. An important implication of WARP is that preference orderings cannot depend upon the addition or subtraction of non-maximal options from the opportunity set, and consequently, the question of which choice is rational cannot depend upon the presence or absence of “irrelevant alternatives.” In addition to being taken as an accurate description of human behavior, WARP is generally invested with normative force: people who violate the condition are said to be acting irrationally.\(^{40}\)

However, WARP has also faced persuasive criticism, most notably from Amartya Sen. In a wide body of work,\(^{41}\) Sen argues that rational decision-makers may violate WARP when their preferences are menu-dependent – that is, when

\(^{40}\) For example, Arrow’s Impossibility Theorem takes independence from irrelevant alternatives as one of the features that a normatively appealing social choice function should satisfy. See Arrow (1951).

\(^{41}\) See e.g., Sen (2002), pp. 165-175.
preference orderings depend upon which other options are also available.

Preferences that are menu dependent can violate WARP because the presence or absence of non-maximal options can affect which option is maximal. If \( x \) is preferred to \( y \) when \( z \) is unavailable, but \( y \) is preferred to \( x \) when \( z \) is available, the preference revealed by the choice from \( \{x,y,z\} \) will not match the preference revealed from the choice between \( \{x,y\} \), thus violating WARP.

Viewed against the backdrop of our previous discussion, we can see that menu-dependent preferences are a special case of the broader problem with identifying consistency in the revealed preferences approach. What the issue boils down to is whether it can be rational for people to have priorities that are contingent upon the other available options. There are at least three cases in which rational actors may have preferences of this type.

First, the available options may derive symbolic meaning from other available options; what the option is depends on the menu of options. For example, suppose that the produce party guest must decide whether to take the last pineapple from a bucket of fruit. Her preferences in this situation may be different from her preferences in a situation with several pineapples left, because the action of taking the pineapple would not then be interpretable as taking the last pineapple. Because taking the last pineapple is symbolically different from taking one of several available pineapples, it may invoke different priorities. Thus the presence or absence of the other pineapples can affect one’s preferences between the pineapple in question and the other types of fruit. Similarly, choosing to fast is not the same as being forced to starve; it is the availability of the option to eat that instills fasting with its symbolic
value. Other examples include being the first to cross a picket line or taking the biggest piece of cake.\textsuperscript{42}

Second, the menu of available options can change agents’ preferences by altering their beliefs about which option will be most consistent with their priorities. To illustrate how non-maximal options can have this epistemic quality, Sen provides the following example: “… Given the choice between having tea at a distant acquaintance’s home (x), and not going there (y), a person who chooses to have tea (x), may nevertheless choose to go away (y), if offered – by that acquaintance – a choice over having tea (x), going away (y), and having some cocaine (z).”\textsuperscript{43} In this example, the availability of the option to use cocaine provided information about the situation and can thus reasonably affect preferences over the other two available options as well.

Finally, the availability of non-maximal options can alter preference orderings by changing the utility of the other options. Recent research conducted by the psychologist Barry Schwartz suggests that the satisfaction people experience after picking a particular option depends on the other options that were considered but passed over. For example, suppose that a student must choose whether to spend her spring break hiking in California or staying with friends in Maine. Going to Maine would be nice because she could spend time with her friends, but California’s warm weather is also appealing. Suppose that all considered, the student prefers hiking in California, even though it means not spending time with her friends. Schwartz’s claim is that in situations like this, the presence of the non-maximal option – going to

\textsuperscript{42} Sen, \textit{Rationality and Freedom}, 130-1.
\textsuperscript{43} Sen, \textit{Rationality and Freedom}, 131.
Maine – can reduce the decision-maker’s satisfaction with her final choice by highlighting foregone benefits. In this way, introducing a new “second-best” choice into one’s opportunity set can decrease the happiness associated with the old best choice, and thus make a third option preferable to both. Through these three mechanisms, rational preferences can be contingent on the availability of so-called irrelevant alternatives.

These cases demonstrate that rational decision-makers may have preferences that are contingent on the menu of available options. When a decision-maker’s preferences are menu-dependent in this way, changing which options are available to her can change the preference that she reveals, thus violating WARP. But because the source of the inconsistency lies in the contingent nature of her underlying (rational) preferences, the deviation from WARP is not evidence of irrationality.

By showing that the standard of rationality informed by the revealed preferences approach fails to avoid reliance on external considerations, we have undermined one of the primary benefits of that approach. But there is a deeper problem with revealed preferences theory that makes it particularly unsuitable for normative analysis. When preferences are defined descriptively – as choices people will make when presented with a decision – then every voluntary decision must necessarily be maximally consistent with the person’s preferences. Thus from a revealed preferences approach, voluntary decisions (that are not inconsistent with previous decisions) are necessarily rational. The problem with the resulting theory of rationality is that it leaves no room to evaluate the quality of a decision; any other

---

choice would have been equally rational because the preferences that determine
erality would have been defined according to that choice instead. Thus the
revealed preference view of rationality misses out on the central function of
instrumental models of rationality. People do have concerns and priorities that they
are trying to bring about, and some ways of acting are more effective for bringing
about those priorities than are others. Defining preferences in a way that ignores
one’s goals leaves us unable to assess the quality of decisions in terms of these
objectives.

A point that deserves emphasis is that referring to a decision as rational in the
sense of revealed preferences does not imply that it is equally rational in the sense of
goal-satisfaction. As we shall see in Chapter 3, it does not follow from the fact that
someone chooses to do something voluntarily that they will also be acting in the way
that best satisfies their priorities. To claim that the two views of rationality are
equivalent is to impose a strong descriptive claim about human behavior: that when
presented with a voluntary choice, people will always choose in the way that best
furthers their priorities. As we shall see, however, deviations from this way of acting
are both theoretically possible and empirically common. As such, a theory of
rationality derived from a revealed preference approach does provide information
about whether the decision is consistent with the decision-maker’s goals.

Before moving on, I want to clarify that I am not claiming that the revealed
preference approach is not useful as a methodological strategy for determining the
content of people’s preferences. Indeed, it may well be the case that focusing on
people’s observed choices is a better source of information about their priorities than
other means of eliciting that information. This information can then be used in positive applications, such as predicting how people will actually choose in future situations. However, there is an important difference between using revealed preferences as an epistemic approach for obtaining information about one’s preferences and defining preferences in terms of those choices. The key question is whether the choices are indicative of preferences or constitutive of preferences. It is important to be explicit about the answer because the different ways of conceptualizing preferences give rise to very different normative interpretations of what it means for a decision to be rational; either rationality indicates no more about a decision than that it is voluntary, or else it involves consistency between the decision and the goals of the decision-maker.

In this section, we have contrasted an instrumental approach to rationality with the revealed preference approach pervasive in economics. We saw that the revealed preferences approach fails to provide a standard of rationality that relies only on internal consistency of choice, and also why it is unsuited for normative analysis. The inadequacies of that approach – the most popular alternative to instrumental models – lends support to a model of rationality oriented around decision-makers’ goals and priorities.

The task that we set out for ourselves in Chapter 1 is almost complete. We have put forth a model of rationality designed to accommodate the broad range of concerns that decision-makers may have, and examined what restrictions are necessary to impose on those concerns if they are to inform rational decision-making. Furthermore, we distinguished our model of rationality from the related theories of
morality and the revealed preferences approach. Armed with the tools developed in this chapter, we can now turn our attention to a class of situations that poses difficult questions for rational analysis: self-conflict. In addition to examining whether such situations can be usefully approached within normative models of decision-making, we will also investigate what rationality prescribes for decision-makers experiencing self-conflict. As we take on these difficult questions, we will be relying on the model of rationality developed in this chapter. Let’s see how far it can take us.
Most decisions that people make involve comparing consequences that occur at different points in time. From choosing whether to spend one’s paycheck immediately or put it into savings, whether to order a greasy hamburger that may impair health later in life, or whether to put off an unpleasant task until tomorrow, people are frequently forced to engage in what economists and decision-theorists refer to as intertemporal choices. Because this type of decision is so pervasive in everyday life, it is crucial that models of decision-making be able to account for intertemporal considerations.

This chapter is divided into four sections. In the first, I describe dynamic conflict and set up the question of whether rationality requires decision-makers to exercise prudence when making intertemporal decisions. In Section II, I take on this question by examining how differing conceptions of personal identity imply different prescriptions for rational choice in situations of dynamic conflict. After concluding that the personal identity approach to answering this question is insufficient, Section III argues that the rationality of prudence derives from the conditional nature of the priorities expected to change. Finally, Section IV applies these results to a separate but related issue concerning the rationality of discounting future consequences.
I. Dynamic Conflict

From the perspective of a normative theory of rationality, a central question is how people should act when making decisions about the future. Answering this question is complicated by a form of self-conflict that can arise when making intertemporal choices. *Dynamic conflict* occurs when a decision-maker expects her future self to develop goals that conflict with the goals she currently has. The possibility of dynamic conflict raises the question of whether rationality requires that decision-makers take their future priorities – priorities that they currently lack but expect to develop – into account when making decisions in the present.

To illustrate dynamic conflict, consider the following scenario: I am deciding whether to learn some basic Italian in preparation for my trip to Italy in six weeks. Although I currently have no interest in learning how to speak Italian and would rather spend the afternoon watching a movie, I expect to develop a strong interest in knowing the language once I arrive in Italy. Is it irrational for me to spend my day going to the movie instead of learning Italian? This decision invokes dynamic conflict because my expected future interest in knowing Italian conflicts with my current preference to watch a movie.

Examples such as this raise the question of whether rationality requires that people take their future goals into account when making decisions about the present.46

---

45 Based on an example from Nagel, 68-9.
46 One could plausibly object to the way in which I am distinguishing motivational influences from current and future priorities. If a decision-maker attaches value to his future priorities when making decisions, the “future priorities” are effectively current priorities already. However, we can accept this argument while recognizing that its force is largely semantic. Someone who prefers this view can think of the term “future priorities” as shorthand for a current priority to further future interests and the term “current priorities” as shorthand for other current priorities.
People who fail to take their future concerns into account when making decisions are often criticized for their short-sightedness, yet most decision-theorists do not consider such behavior to be irrational. For instance, Gary Becker and Kevin Murphy argue that even fully myopic decision-making – in which individuals accord no weight to future concerns – can be consistent with rationality,47 and David Gauthier writes, “Our view is that prudence is rational for those who have a considered preference for being prudent, but not for those who on full reflection do not.”48 In this view, concern for the future is just like any other preference that one might have; decision-making that prioritizes the achievement of present goals at the expense of future ones is no more subject to rational criticism than is a preference for apples rather than oranges. As a result, the exercise of prudence, in which decision-makers attach equal concern to their present and future goals, is rationally acceptable but not rationally required. The consumer who spends today what he will need tomorrow may come to regret his decision, but in this view, his imprudent behavior is not vulnerable to the charge of irrationality.

In contrast to most economists, prominent philosophers have argued that it is irrational for decision-makers to attach less weight to goals that they do not yet have, but expect to develop. John Rawls maintains that “Mere temporal position, or distance from the present, is not a reason for favoring one moment over another.

An insight that emerges out of this objection is that the motivational force of a particular future priority is contingent on the decision-maker’s expectation of developing that priority in the future. For example, my expectation that I will one day want to retire to Florida may prompt me to take actions now to make this possible, but if my expectation changes so that I no longer expect to develop this priority, I will no longer find it a persuasive reason for action.

Future aims may not be discounted solely in virtue of being future…”\textsuperscript{49} Similarly, Jon Elster claims that “It is irrational if one knowingly does what will make one’s life as a whole turn out worse than it could have been.”\textsuperscript{50} For proponents of this prudential view, only decision-making that accords equal weight to present and future concerns can be considered rational.

Both views of rationality have some intuitive appeal. On the one hand, decision-makers’ priorities are by definition the things that they care about. If an individual currently lacks a particular priority, then it follows that she has no rational reason to take that priority into account when making her decisions. On the other hand, a normative conception of rationality is fundamentally concerned with decision quality. If decision-makers lack sufficient prudence, they will consistently regret the decisions that they make. Why should we define rationality in a way that endorses such problematic behavior? To begin to reconcile these conflicting intuitions, we need to examine more closely the features that characterize situations of dynamic conflict.

The complications involved in subjecting intertemporal decisions to rational scrutiny emerge from the fact that people’s goals may change over time. Absent this phenomenon, intertemporal decisions could be analyzed using the same standards of rationality as \textit{intra}-temporal decision-making. That is, the consequences of options in different temporal periods could all be analyzed by determining the extent to which they conform to a decision-maker’s unchanging set of priorities. However, when

priorities do change over time, it becomes less clear which goals should be used as
the basis for assessing rationality.

It is important to distinguish preferences that change over time from
preferences that are constant, but oriented towards the future. Decision-makers may
hold priorities that are future-oriented in the sense that they are concerned with
consequences in later temporal periods. For example, I may be motivated to save
money for the future because I know that doing so will prevent me from worrying
about it in the present. I might also choose to put off eating dessert until after my
meal because I value the positive feelings of anticipation I get from looking forward
to it. In both of these cases, my concern for the future is indirectly derived from my
current priority of doing what makes me feel good now. Other currently held
priorities might provide reasons to be concerned with future consequences directly.
For example, most people who care about preserving the environment would be just
as troubled if they found out that an oil spill would occur in ten years as they would
be from finding out about one in the present. Their priority is timeless in the sense
that their aversion to environmental degradation does not depend on when the
degradation takes place. Current priorities that are oriented toward the future in this
way pose no problem for rationality because the decision-maker can use these
unchanging priorities to determine which option is maximal. It is only when options
invoke priorities that the decision-maker does not currently hold that dynamic conflict
arises.

Additionally, not all forms of priority change produce dynamic conflict.
Dynamic conflict only occurs when the new priorities prescribe a different choice for
a current decision than do the currently accepted priorities.\textsuperscript{51} Many new priorities are not backwards-oriented in this way. For example, I may expect to develop the goal of appearing dignified when I reach the age of fifty, but that goal does not necessarily prescribe that I present the appearance of being dignified before I reach that age. Phrased differently, it is quite possible that upon reaching the age of fifty and acquiring this goal, I will not regret my past undignified behavior. In contrast, the goal of knowing Italian that I expect to develop upon reaching Italy is backwards looking because it can be better served by choosing certain options now – e.g., buying an Italian phrase book – than by others – e.g., watching a movie. In general, priority change only causes dynamic conflict when the new priorities can be better satisfied by options that are worse for achieving the current priorities.

So far we have seen that dynamic conflict occurs when a decision-maker expects her goals to change in ways that prescribe different behavior from what her current goals prescribe. When decision-makers find themselves in these situations, they are forced to consider whether they want to take their future priorities into account. To what extent does one’s expectation of developing a new priority in the future give that person a reason to care about the priority in the present? If the prudential view of rationality is correct, then future concerns deserve the same weight as current ones. If, on the other hand, the arguments for prudence prove unpersuasive, we may conclude that decision-makers can be justified in caring less about future concerns than current ones, and discounting future goals should not be dismissed as irrational.

\textsuperscript{51} Put differently, the preferences derived from the old priorities do not match the preferences derived from the new priorities in such a way that the maximal option is different.
The first section of this chapter clarified the situations that produce dynamic conflict occurs and in doing so, raised the question of whether rationality necessitates prudent decision-making. The next section addresses this question through two approaches. The first approach focuses on personal identity and is concerned with identifying the decision-maker as a unit of analysis. The second issue is whether rational decision-makers can be justified in using temporal position as a basis for attaching importance to their various priorities.

II. Prudence and Personal Identity

To determine whether rationality requires decision-makers to act with prudence, we can review the standards for a decision to be considered rational. Fundamentally, rationality refers to consistency between a person’s choices and her priorities. To determine whether a decision is rational – to assess its consistent with a decision-maker’s priorities – it is first necessary to establish just what the decision-maker’s priorities are. In situations of dynamic conflict, we must first determine whether a decision should be judged on the basis of a person’s current and future concerns, or whether only current concerns should be treated as relevant. At the heart of this issue is the question of personal identity, in particular, whether decision-makers are best conceived of as temporally-unified agents or as momentary selves with momentary interests. Both conceptions have some intuitive appeal. On the one hand, I know that it is me who will suffer if I do not learn Italian today, but on the other hand, why should Future Jacob’s interest in speaking Italian motivate me to action when I most definitely do not share his concern right now?
Derek Parfit (1984) offers a justification for discounting future goals by advancing a conception of personal identity in which only current priorities are constitutive of the agent. He distinguishes two views on identity – the simple view, in which selves are conceived of as constant over time – and the complex view, in which selves exist only in the moment and may not be identical to future versions of themselves. The simple view of identity is the one that is most accepted. Historically, most thinkers have understood a single person at different points in time as being united by some enduring, continuous essence – e.g., a soul – that unites disparate temporal selves into a single person. This view of identity is most closely linked with the prudential view of rationality, in which future goals belong to the decision-maker just as much as current ones.

In contrast to the simple view’s claim that human identity ultimately comes down to a non-reducible relationship linking temporally-separated selves, the complex view of identity holds that personal identity is reducible to a sense of shared psychological connectedness between actors. Although less accepted than the simple view, the complex view of identity is not a recent invention. For example, Hume wondered, “What…gives us so great a propension to ascribe an identity to these successive perceptions, and to suppose ourselves possest of an invariable and uninterrupted existence thro’ the whole course of our lives?”

Parfit too finds the complex view to be the more plausible one. He argues that identity should be thought of as a matter of degree, in which shared experiences, psychological and mental connections, and common values all promote strong links

---

between temporally-separated selves. This continuous conception of identity is bolstered by the observation that the differences within a single person at different points in time may be at least as great as differences between separate people. For example, a heroin addict may feel like an entirely different person when he is satiated than when he is in pursuit of a fix, and vice-versa. The priorities and mental state of the former may be quite different from those of the latter. The complex view of identity allows us to acknowledge this disconnect in a way that the simple view skips over. Similarly, consider the plight of a woman who has been diagnosed with a rare disease that will cause her to become singularly obsessed with collecting pre-colonial Burmese finger puppets.⁵³ A simple view of identity would require her to embrace this future priority as her own and to take steps to ensure its success. In contrast, the complex view allows her to be aware of “her” future goals without identifying them as her own.

Accepting the complex view has straightforward implications for the role of prudence in a theory of rationality. Temporally-distinct selves are not linked together in fundamental ways, but rather are connected only to the extent that certain relationships hold between them.⁵⁴ Consequently, it is perfectly rational for decision-makers to proportion concern for the interests of their later selves on the degree to which they feel connected to those selves. Because our relationship to our future selves is not fundamentally different from our relationship to other people, it is just as

---

⁵³ The example comes from Shane Frederick, “Discounting, Time Preference, and Identity” (Diss., Dept. of Social and Decisional Sciences, Carnegie Mellon University, 1999), 36.
⁵⁴ The relationships that foster identity between selves over time can also hold between spatially separated agents. Parfit’s larger point is that our concern for our own future should often be weaker than our concern for others to whom we feel strongly connected.
rational to discount our future goals as it is to attach less weight to the goals of other
people. Furthermore, because our level of connectedness to future selves diminishes
as we think further into the future, it makes sense that our level of discounting
increases proportionately. As Parfit puts it:

My concern for my future may correspond to the degree of
connectedness between me now and myself in the future … It can be
rational to care less, when one of the grounds for caring will hold to a
lesser degree. Since connectedness is nearly always weaker over
longer periods, I can rationally care less about my further future.\(55\)

One objection to this complex justification for discounting might be that it
leaves us unable to criticize short-sighted behavior. There seems to be something
wrong with undertaking behavior that is predictably regrettable; after all, people
generally try to avoid acting imprudently. However, Parfit’s view of identity seems
to conflict with our intuitions by rationalizing behavior that most people consider
problematic. Does the complex theory of identity render us unable to criticize
imprudent decision-making? Although the arguments presented thus far imply that
imprudent behavior may be rational, the complex view may actually augment our
ability to criticize imprudence by allowing us to draw on ethical considerations. After
all, Parfit’s argument is essentially that people should use the same standard for
apportioning concern to future selves as they use when they apportion concern for
other people. The flip side of this coin is that future selves acquire the same moral
status as other persons. If one subscribes to an ethical theory that supports taking the
interests of other moral actors into account, then by the same logic, decision-makers

are ethically obligated to consider the interests of future versions of themselves.\textsuperscript{56} In other words, imprudent behavior is wrong for the same reasons that any type of selfish behavior is wrong. For example, laws that prohibit children from buying cigarettes can be seen as defending the health interests of their future selves.\textsuperscript{57}

Despite its ability to match most of our intuitions, several observations point toward important problems with the complex view sufficiency as a basis for personal identity. In particular, conceptualizing identity in terms of connectedness – with full connectedness corresponding to complete identity – appears to leave something out of the equation. To illustrate, consider an objection raised by the philosopher Eddy Zemach:

Parfit is forced to say … that if he were to die tomorrow he would find great consolation in the fact that someone who is very much like him will come from Mars to take possession of his house, sleep with his wife, finish and publish the book that he (Parfit) has been working on, and get the total love and devotion of his (Parfit’s) children! If all this is candid … Parfit must be very different than most of us … [If connectedness] is the only item in the new metaphysical framework which roughly corresponds to personal self-sameness in the old metaphysical framework, then it takes on … the function of the old concept in the old framework. This is a methodological mistake and psychologically absurd.\textsuperscript{58}

Because the central tenet of the complex view is that identity is reducible to physical and mental connectedness, a decision-maker who accepts Parfit’s claim and allocates concern based on these factors would have to value the goals of his perfect

\textsuperscript{56} This point comes from Frederick, 43. Note, however, that such criticisms must proceed on ethical rather than rational grounds.

\textsuperscript{57} More broadly, this view implies that government policies that seem paternalistic may be justifiable on grounds that they protect future citizens – or rather, future versions of current citizens.

replica just as much as he values his own concerns. To the extent that such a claim
seems absurd, the complex view must be insufficient as a foundation for personal
identity. However, thought experiments of this type can only go so far in responding
to Parfit’s arguments. After all, Parfit acknowledges that the simple view of identity
is the one that has been traditionally accepted, and as a result, the intuitions that
inform these thought experiments may stem from a false metaphysical belief.

Christine Korsgaard (1989) defends our intuitions by developing a deeper
critique of the complex view of identity. Although she accepts Parfit’s claim that
“there is no deep sense in which I am identical to the subject of experiences who will
occupy my body in the future,” she argues that Parfit fails to appreciate the
significance of agency in his discussion of identity. Agency, she contends, is what
unifies temporally distinct selves into a single person. Later selves may be quite
different from earlier selves and may have quite different priorities, but in general,
those differences are created by the agents themselves. That a common agency lies at
the root of these changes is at the core of Korsgaard’s theory of identity: “Where I
change myself, the sort of continuity needed for identity may be preserved, even if I
become very different.” In other words, although a decision-maker might expect
her future self to have priorities quite different from her current ones, she should also
recognize that it is she herself who is the author of those differences. Consequently,
decision-makers are linked to their future selves in a fundamental way that the
complex view of identity skips over: we are the authors of our future selves. This

60 Korsgaard, 123. Elster makes a similar point: “I believe … that a person who takes his future states as given, rather than something to be created, is fundamentally irrational.” Elster, *Rational Choice*, 11.
insight helps us understand why our concern for ourselves is not identical to our concern for others who are quite similar to us, but whose choices are not tied to our agency. Similarly, it helps explain why the heroin addict and the women infected with the puppet disease may not identify with their future selves, because the sources of the change are outside their control.

Although persuasive, Korsgaard’s argument is incomplete as a justification for prudential rationality. Even after recognizing future priorities to be their own, and even after identifying with them as such, why are decision-makers rationally obligated to accord them equal status with their other, current priorities? After all, most people do not attach equal importance to all of their priorities. For example, almost everyone has, at least to some extent, a priority to act morally as well as a priority to pursue happiness. However, some people attach more weight to the former whereas others attach more weight to the latter, and neither of these positions is obviously irrational. Could a decision-maker be similarly rational by according more weight to her current priorities than to her future ones? Our question now becomes whether it can be rational to use temporal position as a basis for choosing how to weight conflicting priorities.

Up to this point, our goal of developing a conception of rationality capable of accommodating a wide range of reasonable goals has prevented us from imposing requirements on what priorities may be rationally acceptable. At this stage, however, it seems reasonable to introduce a relatively innocuous condition along such lines. In particular, we can claim that for a priority to be rational, it must be able to withstand reflective scrutiny on the part of the decision-maker who holds it. This condition
seems plausible; unless priorities appear to decision-makers as persuasive reasons for acting, it is hard to imagine how they could carry any motivational force. By the same logic, a similar standard should apply to weighting schemes between priorities.

Although satisfaction of this condition for a particular weighting scheme is ultimately an empirical issue, it seems all but certain that some methods of allocating concern among priorities would fail even this minimum standard. For example, consider a weighting scheme in which priority conflicts were resolved on the basis of the alphabetical order of the second letter in the priority’s name. For decision-makers who determined such a method to be unsatisfactory, employing this weighting scheme would be irrational. Although it is not a priori impossible that someone could find this standard to be an acceptable method for making the decisions in her life, it would probably seem unreasonable to most people and would therefore fail their reflective scrutiny tests.

We can take a similar approach when trying to determine whether a time-sensitive weighting scheme for choosing between priorities is inherently irrational. The question we must answer is whether a priority – or more specifically, a standard for allocating concern between priorities – that takes a priority’s temporal position into account could be accepted by a decision-maker upon reflection. If we are convinced that this standard would be rejected by decision-makers who reflect upon it, then we could conclude that it is irrational in the same sense that we judged the alphabetical weighting scheme to be irrational in the previous example.

---

61 This requirement is not entirely new. We implicitly referred to a similar standard in Chapter 1 in order to show that merely choosing to adopt a new priority is generally not an effective strategy for satisfying second-order preferences.
Let us consider why a time-sensitive weighting scheme might be judged problematic. The main concern is that the temporal position of the decision-maker seems like an arbitrary basis for allotting concern between priorities. Accepting as rational a current priority that is expected to change later is to acknowledge that the priority is only persuasive if one happens to reflect on it during a particular stretch of time. Judging such a priority to be rational requires a willingness to believe that one’s temporal position can be a relevant consideration when deciding what values and goals are worthwhile to pursue. In other words, decision-makers who embrace time-sensitive weighting schemes must not be bothered by the fact that their judgment about which priorities are most important would be different if they were making the judgment at a different point in time. For example, someone who is rationally motivated by a priority that attaches more weight to happiness in $t_0$ than to happiness in $t_1$ would need to accept the fact that this priority would lose its appeal if it were reflected upon at $t_1$. Strong arguments about the problems with such a judgment have been articulated by thinkers ranging from Sidgwick to Nagel. Assuming that we are sufficiently persuaded by their case, we could argue that since willingness to accept temporal position is based on mistaken reasoning, decision-makers who adhere to time-sensitive priorities must be acting irrationally.

We should not take this path. To argue that such priorities are inherently irrational would be to claim that anyone who gives the issue considered judgment would come to the conclusion that temporal position is an unacceptable basis for a weighting scheme. However, as the work of scholars such as Parfit demonstrates,

---

62 See for example Sidgwick (1874) and Nagel (1970).
reasonable people can examine the issue and conclude that temporal position can be relevant to a priority’s motivational force. The issue is not whether Parfit is correct in his view of identity – we have already seen why his view is mistaken – but whether people who agree with Parfit’s conclusion are irrational for holding that view of identity and employing the associated weighting scheme.

Resolving issues of this sort boils down to deciding how accommodating we believe that a theory of rational belief should be. If a particular belief – either empirical or in this case metaphysical – is plausible but incorrect, should people whose priorities are informed by that belief be considered irrational? In most cases, I believe the answer should be no. If we want our model of rationality to be able to accommodate the broad range of priorities that sensible people may hold, then we need to allow some room for what Rawls refers to as reasonable disagreement. Although a more in-depth discussion of this issue would be necessary for extending our model of rationality to cover beliefs, such a project is outside the scope of this thesis. It seems likely, however, that a theory of rational belief grounded upon the above principles would not condemn as irrational decision-makers whose weighting scheme depended upon temporal position.63

This section dealt with the question of how to define the decision-maker as a unit of analysis, and in particular, whether rationality should be assessed on the basis

---

63 A different problem may be that Nagel’s argument is itself flawed. His claim that valid reasons for action cannot depend upon who is considering them (agent-neutrality) or the circumstances in which they are considered (e.g., time-neutrality) conflicts with many of the priorities that people view as legitimate. In particular, accepting Nagel’s claim would mean treating as irrational concerns whose persuasiveness derives from the circumstances of the actor. For example, most people believe that it can be rational to buy birthday presents for their own child, despite the fact that they would lack interest in the particular child if they found themselves in different circumstances, e.g., the parents of a different child.
future priorities in addition to current ones. Although we rejected Parfit’s complex view of identity – which would have excluded future priorities from consideration – we also saw that reasonable people may choose to attach more importance to current priorities than to future ones. The appeal of such a weighting scheme stems from beliefs about identity that are misguided, but not irrational. As a result, accepting a Korsgaardian view of identity is not sufficient to demonstrate that rationality mandates prudence on the part of decision-makers. Even when decision-makers recognize future priorities to be their own, we have not shown that it is irrational for them to accord less weight to their future priorities than to their current ones. In the next section of this chapter, I attempt to extend Korsgaard’s important insights into a stronger argument on behalf of prudent behavior.

**III. Conditional Priorities**

In Section II, we drew upon Korsgaard’s work to reject the complex view of identity and the justification for imprudent decision-making that went along with it. However, we saw that even after accepting a Korsgaardian approach to identity, in which decision-makers are tightly linked to their future goals, we were still unable to conclude that imprudent behavior is rationally prohibited. However, Korsgaard’s insights about the relationship between people’s current and futures selves can point us in the right direction.

A good starting point is her observation that although decision-makers’ future priorities may be very different from their current ones, in general those changes are of their own making; for the most part at least, people’s goals only change when they
choose to change them. We can make headway on the question of how decision-makers should take future priorities into account by considering the reasons that priorities may differ between present and future versions of the same person. As we shall see, the rational way to respond to anticipated priority changes depends upon why we expect those changes to come about.

Why might a decision-maker’s priorities change over time? Broadly speaking, we can divide the sources of priority change into two categories: those that are acceptable to decision-makers and those that are unacceptable. Sources of priority change may be unacceptable for at least two reasons. First, priorities might change for reasons that are not in the decision-maker’s control. For example, the rare disease that causes victims to become obsessed with Burmese finger puppets would fall into the “unacceptable” category because it forces people to have goals that they do not choose to accept. Similarly, a recovering addict may recognize that his chemical dependency on heroin will cause his priorities to change, but he may view this change as unacceptable because he lacks control over the alteration. The second reason that priority change may be unacceptable is when it is caused by deterioration in the quality of the decision-maker’s judgment. For example, most people’s priorities reflect what they believe to be important. If a decision-maker anticipates that his future judgment will be of lesser quality than his current beliefs, the new priorities that result from the inferior beliefs will also be of lesser quality.64

---

64 The opposite is also true. If we attribute our changing priority to an anticipated improvement in the quality of our judgment, e.g., that we expect to become wiser, we should give full weight to our future priority. By acknowledging that we expect our improved judgment to yield a different set of priorities from our current one, we have effectively already changed our mind and accepted the greater validity of the future priority.
When the source of priority change is unacceptable, imprudent behavior is eminently rational. After all, our rejection of Parfit’s view of identity was predicated on the assumption that decision-makers are responsible for the differences between their current and future selves. However, when decision-makers perceive the sources of change to be unacceptable, Korsgaard’s argument no longer holds. As a result, decision-makers are no more rationally obligated to identify with their own future priorities than they are to identify with the priorities of other people. When dynamic conflict is caused by priority change of this type, the prescriptions for rational behavior are straightforward.

We can contrast this state of affairs with the situations that occur when decision-makers expect to undergo priority change, but determine that the source of this change is acceptable. If an initial, acceptable set of priorities changes in an acceptable way, then the resulting set of priorities must also be acceptable. In other words, decision-makers who start from a persuasive set of priorities and expect those priorities to change for reasons they deem to be acceptable will find the resulting set of priorities equally compelling. The key point here is that as long as a decision-maker expects her priorities to change for acceptable reasons, she must simultaneously acknowledge the validity of both her current and future concerns.

Furthermore, because the current and future priorities do not hold at the same time – the current priority changes into the future one – the future priority must be judged to be valid for the future but not for the present and the current priority must be judged to be valid for the present but not for the future.65 How can we reconcile

---

65 By valid, I mean the decision-maker finds the priority to be a compelling reason for action.
these judgments? For the present concern to be valid now, but not in the future, it
must be the case that its validity is conditional on some circumstance that changes
between the present and the future. Similarly, for the future concern to be valid in the
future but not in the present, its validity must also be predicated on some changing
condition, one that holds in the future but not in the present. Finally, because the
present and future concerns do not hold simultaneously – the former changes into the
latter – we can describe them both as manifestations of a single, conditional priority.

An example can help illustrate this concept. Suppose that I am a member of a
family that has a very rigid tradition, in which the oldest living member of the family
is expected to become an expert on the family’s entire genealogical history. I may
currently have no interest in learning about my family’s genealogy, yet
simultaneously acknowledge that if I become the oldest living family member, I will
find learning about our genealogy to be a worthwhile priority. The fact that I accept
the validity of both my current and future judgment on the merits of embracing the
priority implies that my current priority is actually conditional on whether or not I am
the oldest living member of my family: I do not care about genealogy when I am not
the oldest living family member and I do care about genealogy when I am the oldest
living family member.

In the preceding example, my priority was conditional on an external factor, a
circumstance that could trigger priority change by changing over time. Other
priorities might be conditional on a factor that is certain to change over time, namely
time itself. Suppose that my priorities are such that I accord more importance to
present (t₁) happiness than to happiness I expect to experience next week (t₂).
However, I fully expect my priorities to change over time so that by next week (t₂), I will attach more importance to happiness experienced then (in t₂) than to happiness experienced today (in t₁). Furthermore, I perceive my present concern to be a compelling reason for action and I also accept that my future concern will be equally compelling next week when I choose to adopt it. As before, we can combine these two concerns into a single priority whose content is conditional on an external circumstance: the temporal position of the decision-maker. This week, my priority tells me to emphasize happiness now, whereas next week, the same priority will instruct me to maximize happiness then.

To summarize the argument so far, I have claimed that when decision-makers accept the validity of both present and future concerns, it must be the case that the persuasiveness of their current priority is actually conditional on some external circumstance. Because the same arguments holds for future priorities that are not accepted as valid in the present, we can conclude that decision-makers who undergo priority change for acceptable reasons are actually being motivated by a single, conditional priority. The final step to our argument is showing that when priorities are conditional in this way, rationality dictates that decision-makers exercise prudence when making their decisions.

To determine the rational way to achieve conditional priorities, it will be helpful to treat the situation with some abstraction. Suppose that a decision-maker is currently motivated by goal A, but expects her priorities to change so that in the future she is instead motivated by goal B. Assuming that she sees the source of these changes as acceptable, we saw that the motivational force of both A and B are
contingent upon some external condition X. As a result, we can combine A and B into the conditional priority C, which instructs the decision-maker to promote A when X is true and to promote B when X is false. Our concern now is how best to promote C.

As we saw in Chapter 1, unconditional priorities are best promoted by choosing the option whose consequences are most consistent with the priority’s aim. For instance, if our aim is to minimize environmental degradation, then we should choose the option that causes the least amount of damage to the environment. The case is similar when priorities are conditional. To promote C, we should pick the option that best promotes its (conditional) aim. In other words, we should pick the option that maximizes the satisfaction of A when X is true and the satisfaction of B when X is false.

A simple model is well suited for analyzing this type of situation. For simplicity, we can restrict our focus to two periods, with X being true in the first period and false in the second. To test whether conditional priorities are best served by prudent decision-making, we can compare two options, one prudent and one focused exclusively on current goals. Both options have two consequences, one occurring in each time period. Option 1, the imprudent choice, is maximally consistent with the decision-maker’s current concerns, in this case goal A. As such, its effects are maximally consistent with goal A in both the first and second periods. Option 2, the prudent option, takes future goals into account. It results in the consequence most consistent with option A in the first period and the consequence most consistent with option B in the second period.
It comes as no surprise that when we compare the two options, the prudent option does a better job achieving the decision-maker’s priorities. Whereas Option 1 satisfies C in the first period only, Option 2 satisfies C in both periods (see Table 1 below). From this example, we see that decision-makers will be better served by choosing options whose intertemporal consequences will be consistent with the decision-maker’s concerns in the time period in which the effects occur. We are finally in a position to answer the central question of this chapter: rationality dictates that decision-makers take future concerns into account when experiencing dynamic conflict in the present. More specifically, if a decision-maker determines her current priority to be conditional on an external circumstance, then she is rationally obligated to take into account changes to that priority from expected changes in the underlying circumstance.

**Table 1 – Prudence and Conditional Priorities**

<table>
<thead>
<tr>
<th></th>
<th>Option 1</th>
<th>Option 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Period 1</strong></td>
<td>Satisfies A (and C)</td>
<td>Satisfies A (and C)</td>
</tr>
<tr>
<td><strong>Period 2</strong></td>
<td>Satisfies A (but not C)</td>
<td>Satisfies B (and C)</td>
</tr>
</tbody>
</table>

We can make use of the above reasoning to respond to an objection that Parfit raises against prudential views of rationality. He claims that if temporal position is not relevant for determining whether to take a priority into account, then it must also be the case that *past* goals that the decision-maker no longer cares about are just as important as future goals that the decision-maker expects to develop. Is this claim true? When deciding where to apply for graduate school, does a college student need to take into account her heartfelt goal at age eight to one day become an astronaut? If
not, then how can she justify taking into account concerns that she expects to develop in the future?

To respond to Parfit’s objection, we can again distinguish between the reasons that priorities might change over time. In the simple case where the change is due to an improvement in the quality of the decision-maker’s judgment over time, the earlier priority can be reasonably dismissed as being based on inferior judgment. For example, she may have lost her desire to become an astronaut as her judgment about worthwhile life pursuits and what professions she is best suited for evolved over time. Conversely, if she recognizes that the change in goals stemmed from a force outside of her control, she may want to revert back to her earlier priority, depending on whether she views the external force as acceptable.

We can now turn to cases in which the decision-maker views both her current and past concerns as equally acceptable for their respective temporal positions. As we have seen, when priority change occurs in this way, it must be the case that the decision-maker’s overriding priority is conditional on an external factor. We have seen that acting rationally in such situations entails taking a concern into account only to the extent that its antecedent condition is satisfied. As such, one may be rationally obligated to take past concerns into account when making current decisions, but only to the extent that the antecedent conditions to those concerns are expected to be satisfied in the future periods in which the decision’s consequences will occur. In other words, decision-makers should only attach importance to past concerns when they expect those concerns to be theirs again at some point in the future. In such
cases, however, the past concern’s importance derives from its status as a future goal, not as a past one.

This section has argued for a prudential view of rationality by considering the reasons that decision-makers may experience priority change. I claimed that when decision-makers view such changes as acceptable and recognize the validity of both their present and future concerns, it must be the case that both concerns are actually manifestations of an overarching, conditional priority. And because conditional priorities are best satisfied by taking into account each conditional prescription based on the likelihood of its antecedent condition obtaining, rational decision-makers should take their future goals into account when making present decisions.\textsuperscript{66} Rationality demands prudential decision-making. In the final section of this chapter, I explore what implications this conclusion has for a popular approach to analyzing intertemporal decision-making.

**IV. Discount Rates and Pure Time Preference**

At the beginning of this chapter, we noted that people are frequently forced to choose between options whose consequences occur at different points in time. Models of intertemporal decision-making have long recognized the descriptive fact that people tend to weight imminent consequences more heavily than distant ones when making

\textsuperscript{66} Although we have seen that people’s priorities can change based on changing circumstances, our discussion thus far has treated the changing circumstances as exogenous. However, it is often the case that decision-makers have some control over the circumstances that their priorities are contingent upon. For example, although I may accept the priority of wanting to know Italian as a compelling motivation when I am in Italy, I also have control over whether I travel to Italy in the first place. To the extent that I am in a position to control which priorities I develop, I may find myself confronted with a decision with options that differently determine my future self’s priorities. Resolving such questions is thus tied to higher-order preferences in a basic way.
decisions. When people act in this way, economists say they are discounting the future. In this section, I explore the connection between this type of discounting behavior and the form of dynamic conflict examined in the first part of this chapter. As we shall see, the question of whether it is rational to discount future consequences is related to, but distinct from the question of whether it is rational to discount future goals. In particular, we can ask whether rationality prohibits decision-makers from treating future consequences as less important than imminent ones.

We can begin to address this question by disentangling some of the reasons that people may choose to prioritize imminent consequences over future ones. Decision-makers may choose to discount the future for several reasons. First, they may take into account the additional uncertainty associated with future outcomes. One of the reasons I prefer a gift of $100 today to the same gift in ten years is because I know that there is some risk of my dying in the intervening period, an event that would render me incapable of enjoying the gift. Less extreme possibilities also merit consideration. If I expect to have more money in ten years than today, diminishing marginal utility of income implies that I will value the absolute increase in wealth more today than in ten years. The value of the gift may also change over the intervening period. A bout of hyperinflation may render $100 of today’s currency worthless in ten years. Finally, my future goals are themselves subject to some uncertainty. In ten years I may have renounced the material pleasures purchasable with money and not value the gift at all. Although such changes may be less likely in some cases than in others, some adjustment for uncertainty seems advisable in virtually all intertemporal decisions.
Two other factors that are pertinent to intertemporal decisions are the roles played by anticipation and memory. In the days before a pleasurable event, most people derive anticipatory pleasure, and similarly, in the days following the event, they may derive pleasure from the event’s memory or regret from the fact that the event is over. For unpleasant events the situation is reversed, with people experiencing dread from impending pain or discomfort. The periods after the unpleasant event may be characterized by unpleasant memories (even trauma in extreme cases) or by relief that the unpleasant event is done with. These factors can help explain the finding that while people tend to be impatient with respect to pleasurable outcomes, they also tend to get unpleasant outcomes over with as quickly as possible, for example to avoid negative anticipation.67

A final factor that can prompt decision-makers to prioritize present benefits over future ones stems from the fact that the passage of time may render a given outcome in the present different from an otherwise equivalent outcome in the future. For example, money received today can be invested and accrue interest. A gift of $100 today, invested at a ten percent annual interest rate is equal to $110 in one year. Because of this interest rate effect, rational decision-makers should prefer to receive monetary benefits sooner rather than later. A similar effect may characterize non-monetary considerations as well. If I receive a chocolate cake, I may enjoy the anticipation gained by putting off its consumption, but as the cake gets older and its freshness diminishes, I need to balance the pleasure from anticipation against the fading benefits from consumption.

All of the reasons for discounting discussed thus far are motivated by a recognition that temporally far off consequences may be differently related to the satisfaction of one’s priorities than temporally-near consequences. To the extent that rationality concerns the degree to which options are consistent with priorities, decision-makers are perfectly justified in taking these considerations into account. For example, if my objective is to maximize the amount of money I can earn, it would be irrational for me to neglect the role of inflation or to treat future payments as equivalent to present ones. Thus we see that at least some forms of discounting are consistent with rational decision-making.

On the other hand, not all forms of discounting are so clearly consistent with rational decision-making. Apart from the considerations discussed thus far, a decision-maker may also choose to discount the future per se, that is, to attach less importance to future consequences for no other reason than that they occur in the future. When decision-makers are motivated to discount in this way, we say that they have a pure time preference.

Is it irrational for decision-makers to discount based on a pure time preference? People develop a pure time preference when one of their priorities attaches more weight to consequences that occur in the present than to those that occur in the future. In contrast, the other reasons for discounting that we have discussed so far come from priorities that are concerned with time only indirectly. For example, future monetary prizes are only discounted over current ones because of factors such as the foregone benefits of investment and the risk of inflation; the temporal position of the consequence is only indirectly relevant to the satisfaction of
the priority. In contrast, if one of my priorities was to maximize immediate
happiness, pursuing that priority would entail my favoring options that bring about
happiness quickly over options that bring about happiness with a delay. In that case,
the temporal position of the consequence would be directly relevant to the priority. In
order to conclude that individuals who have pure time preferences are irrational, we
would need to show that priorities that are directly concerned with temporal position
are intrinsically irrational.

Although demonstrating such a claim is a formidable task, we can draw on our
results from last section. It turns out that such time-oriented priorities are irrational
only when they are expected to be temporary, that is, when following them would
contradict the dictates of prudence. To demonstrate this claim, we can start by
considering a situation that does not invoke dynamic conflict. Recall that dynamic
conflict only occurs when decision-makers expect to undergo a change in priorities.
Therefore, a situation in which individuals have temporally-oriented priorities that are
expected to remain constant may exhibit pure time preferences without being
imprudent. For example, people may have a constant set of priorities that attach
different degrees of importance to different stages of their life. Consider a person
who decides that he should value happiness in college more than happiness at other
points in his life, and to achieve this end, he prioritizes his college years when making
intertemporal decisions. As long as this person does not believe that the
persuasiveness of his priority is conditional on the time in which it is considered,
discounting consequences that occur in non-college years would be perfectly rational.
Indeed, while this person is in college, it will be rational for him to discount
consequences that occur after college relative to more immediate consequences that occur while he is still in college. Because his discounting is based purely on the temporal position of the consequences, it reflects a pure time preference. So although this example is somewhat unrealistic, it demonstrates that it is theoretically possible for pure time preference discounting to be consistent with rational behavior.

On the other hand, and as the implausibility of the above example suggests, such situations are few and far between. For the most part, priorities that attach more weight to a particular stretch of time are not constant over time. We can draw a distinction between priorities that are concerned with a fixed period in time, such as one’s college years, and priorities whose temporal focus is relative to the temporal position of the decision-maker. For example, a priority for imminent happiness is temporally-relative in the sense that the period of time that is focused on shifts as the decision-maker moves through time – the stretch of time I am most concerned with today is different than the stretch of time I am most concerned with tomorrow. Because the time period that these relative time-oriented priorities are concerned with changes as the temporal position of the decision-maker changes, they are conditional on the temporal position of the decision-maker. Therefore, making use of last section’s result, we can conclude that the rational way to achieve such priorities is to act with prudence. Discounting future consequences based on temporally-relative priorities neglects the fact that one’s future priorities will be concerned with those discounted consequences. As long as decision-makers expect the priority motivating their discounting to change in a way that attaches more value to the future
consequences, discounting future consequences on the basis of one’s current priorities is imprudent and therefore irrational.

The arguments in this section pose a major challenge to the agnostic way that discounting is treated in most normative models of decision-making. Although some motivations for discounting the future, such as adjusting for uncertainty, are clearly rational, discounting future consequences just because they lie in the future is generally imprudent. Discounting based on pure time preferences is only rational when the priorities motivating the discounting are both unchanging and focused on a specific stretch of time. Because such priorities are few and far between, however, pure time preferences are almost always irrational.

This conclusion has important practical ramifications for the design of policy. A central question – and a constant source of controversy – in cost-benefit analyses of potential policies is what discount rate to employ when assessing intertemporal consequences. For example, the contemporary debate over the costs of global warming hinges upon what rate is used to discount effects that are expected to occur in the future. Although long-term issues such as global warming invoke separate concerns such as inter-generational justice and a government’s obligation to future citizens, the issues discussed in this chapter have particular relevance for policies whose consequences occur within a smaller temporal window.

One method for obtaining the appropriate discount rate is to measure and then adopt the discount rates employed by citizens who are making their own decisions. However, this approach is problematic if there is reason to suspect that individuals’ intertemporal decisions are often made irrationally. A different approach is suggested
by an understanding of government in which the goal of policy is the satisfaction of citizens’ preferences. In this perspective, the optimal discount rate is the one that will deliver policies most consistent with the preferences of citizens. If it is irrational for individual decision-makers to employ a pure time preference when making their own decisions, it may be similarly irrational for policymakers to make assessments based on a similar concern for the present over the future. In other words, the fact that individuals can best satisfy their own priorities by avoiding a pure time preference suggests that policymakers seeking to satisfy the goals of citizens would be similarly well advised to avoid discounting the future per se. Although the differences between individuals and governments imply that much more consideration of these issues is necessary before drawing definitive conclusions, our results suggest that policymakers need to carefully consider which factors are valid for informing their selection of a discount rate.

This chapter has argued that an adequate conception of rationality requires decision-makers to take into account goals that they expect to develop in the future. We pursued two lines of argument in reaching this conclusion. First, we examined the relation between prudence and personal identity and saw how Parfit’s reductionist view could justify imprudent behavior. Although we concluded that his conception of identity was flawed, we saw that embracing a rival view of identity in which decision-makers are temporally unified was not enough to demonstrate the rational necessity of prudence. In Section III, we took a different approach and argued that when dynamic conflict arises from “acceptable” priority changes, the decision-maker’s underlying priorities must be conditional, and therefore, the rational way to
advance them is through prudent decision-making. Section IV then considered the implications of this result for the discounted utility model of intertemporal decision-making.

An important theme that emerges from this discussion is that the rational way to respond to dynamic conflict depends crucially upon how much confidence decision-makers place in their future selves. When a decision-maker places confidence in her future self’s judgment that a priority is a compelling reason for action, it is rationally necessary for her to take that priority into account when making decisions in the present. In the next chapter, we move on from considering intertemporal forms of self-conflict to cases in which self-conflict divides a decision-maker within a single temporal period. Unlike the intertemporal case examined here, however, self-conflict that occurs in a single period is generally characterized by a disparity in judgment quality that makes it easier to evaluate which side of the conflict is rational.
In Chapter 2, we examined a form of self-conflict that can occur between a person’s current and future interests. In that case, we found that it was necessary to take both sides of the conflict – that is, both current and future priorities – into account when assessing the rationality of a decision. Furthermore, we saw that this result was contingent upon both the current and future interests being authentically the decision-maker’s own, in some fundamental sense. In this chapter, we will show that the same insight applies to a different form of self-conflict, one whose analysis is complicated by the fact it occurs within a single decision-maker at a single point in time. To see what rationality requires in such cases of intra-temporal conflict, we will first need to do some work expanding the simple model of decision-making that has influence our analysis thus far. In doing so, we will discover an important mechanism through which normative and descriptive models of decision-making may diverge.

I. Can Weakness of Will Exist?

Traditional models of decision-making are not useful for shedding light on how people experience and respond to problems of self control. In particular, the standard models used in the social sciences to explain human behavior leave no room for weakness of will, the phenomenon that occurs when a decision-maker voluntarily and intentionally decides on an option that she judges to be less consistent with her priorities than another option that is also available. In other words, weakness of will entails decision-makers knowingly and voluntarily making decisions that they believe
to be worse than other decisions they could make. Before deciding whether we want

to analyze such behavior within the normative conception of rationality developed

thus far, we need to ask whether weakness of will is possible in the first place.

A strong case can be made that weakness of will is self-contradictory and

therefore impossible. When a person chooses from a menu of options voluntarily, she

only decides upon an option if she wants to decide upon that option. Similarly, for a

choice to be intentional, the option that a person ends up picking must be the option

that she decided to pick. From these two definitions it follows that when a choice is

both voluntary and intentional, the option that gets picked must be the option that the

decision-maker wanted to pick. In other words, it is impossible for people to

voluntarily and intentionally choose options that they do not want to choose;

otherwise, the choice would not have been voluntary and intentional. When

approached this way, weakness of will appears self-contradictory because it implies

that a decision-maker wants to choose a particular option while at the same time not

wanting to choose that option.

Before responding to this argument, it will be helpful to consider its

implications for the normative analysis of decision-making. We can think of people

deviating from rationality in at least two ways: they may form irrational beliefs

(epistemic irrationality) or they may make irrational choices given the beliefs that

they have (choice irrationality). The arguments against weakness of will effectively

rule out the latter form of irrational behavior. If a person’s decisions necessarily

reflect her wants, then it must be the case that her voluntary decisions are maximally

consistent with her preferences. That is, if an un-preferred option is freely chosen
over a preferred option, then the un-preferred option must actually have been preferred; otherwise the decision-maker would not have voluntarily chosen it. As such, choice irrationality can only exist to the extent that weakness of will is possible.

This conclusion has important implications for everyday decisions. Consider the following scenario:

**The Vegas Trip**

Upon receiving an unexpectedly large bonus from work one Friday, you and your friend Steve decide to go gambling in Vegas for the weekend. Before entering the casino, Steve confides in you that he has been saving up for an important purchase but has had trouble gambling responsibly in the past. To ensure that he is not tempted to overspend, he asks you to hold on to all of his money, except for a reasonable amount that he plans to gamble with. Because Steve knows that he will be tempted to overspend once he begins to gamble, he asks that you not give his wallet back to him until the end of the day when he leaves the casino. Being the good friend that you are, you agree to this request, and head off to play some blackjack.

After an hour, things are going well for you at the blackjack table. Unfortunately, things do not look like they are going as well for Steve. Indeed, a short time later, he walks over and demands his wallet back. Noting your hesitation, Steve acknowledges, “I know I asked you to hold on to it for me, but I’ve changed my mind and want it back!” You are now faced with a dilemma: do you give the wallet back to Steve (it is his, after all) or do you honor your original promise?

Most people, I think, would recognize the better choice in this example to be holding on to the wallet. Steve recognized that he would be tempted to overspend and took steps to protect himself. However, the rationale for holding on to the wallet is less persuasive if one does not believe that weakness of will can exist in the first place. The impossibility of choice irrationality would imply that Steve’s voluntary decisions necessarily reflect his priorities. After all, Steve is not being forced to
spend the extra money. He will only choose to risk his savings if that is what he wants to do.\textsuperscript{68}

On the other hand, both introspection and observation of everyday decision-making suggest that weakness of will is far from impossible. People frequently make decisions they know they will regret: we choose to lie in bed when we care more about arriving at work on time; we watch bad television and eat unhealthy foods despite judging it better to spend our afternoon at the gym; in general, we give in to temptation while wishing that we were not doing so. As one recent article put it: “Casual observation, introspection, millennia of folk wisdom, and a mass of psychological research all suggest that people … would ‘like’ to behave in one manner, but instead ‘choose’ to behave in another.”\textsuperscript{69}

Thus we see that whether or not our model denies the possibility of weakness of will, people see the phenomenon as a real problem in their lives and consider self control issues to be an important aspect of decision-making. Consequently, if a particular understanding of decision-making rules out such behavior as impossible, this may be a failure on the part of the model rather than evidence that weakness of will does not exist. After all, the primary purpose of developing a normative model of decision-making is to be able to evaluate the quality of decisions. If we conclude that weakness of will concerns can be relevant to decision quality, then we need to extend our model so that it can take those considerations into account. As a first step

\textsuperscript{68} An alternate justification for keeping the wallet may be a suspicion that Steve is experiencing epistemic irrationality and misperceives the odds of winning. We can rule out this possibility by supposing that he successfully passes some test that ensures he understands the likely consequences of his actions.\textsuperscript{69} Ted O’Donoghue, and Matthew Rabin, “Choice and Procrastination,” Quarterly Journal of Economics 116.1 (February 2000), 1.
in that direction, I now attempt to show how weakness of will can be possible in the first place.

Donald Davidson (1969) formalizes the “weakness of will is impossible” argument by advancing the following two propositions:

**P1.** If an agent wants to do x more than he wants to do y and he believes himself free to do either x or y, then he will intentionally do x if he does either x or y intentionally.

**P2.** If an agent judges that it would be better to do x than to do y, then he wants to do x more than he wants to do y.\(^\text{70}\)

If we were to accept both of these propositions, then it would appear that weakness of will is indeed impossible and attempts to model it are fundamentally misconceived. However, the two propositions contain an ambiguity in their use of the word “want.” There is one sense in which “want” refers to a motivational state – to one’s inclination – but there is a distinct sense in which the word refers to one’s priorities and considered judgment. P1 invokes the former meaning whereas P2 invokes the latter. This double usage is misleading because it conflates two distinct concepts through a coincidence of language. Although they often happen to coincide, the link between the two meanings is not unbreakable: the option that I judge to be most consistent with my priorities may not be the option that I am most inclined to select when making a decision. One is a purely cognitive act of judgment whereas the other is primarily concerned with action. There is no deep sense in which these two concepts are fundamentally identical.

To illustrate the distinction, suppose that I must choose between walking home from work and taking the bus. After considering the decision, I determine that the option most consistent with my priorities is to walk home. Although I could save some time by taking the bus, I want to save money on the bus fare and I know that the exercise from the walk will do me good. However, there is nothing about reaching this conclusion that forces me to actually choose to walk home. I can voluntarily get on the bus, even while believing that walking would be more consistent with the goals I care about achieving. In other words, I can still “want” to take the bus despite judging that doing so would be worse than walking.

Some may respond to this example by arguing that the less-preferred action – getting on the bus – must actually be more preferred. Proponents of this view may claim that choosing to get on the bus constitutes prima facie evidence that the concerns motivating me to walk home were weaker because my ultimate, “net want,” was to take the bus. This claim is predicated on the same assumption that motivates the revealed preference approach described earlier, that people’s actions necessarily reflect their true objectives. However, imposing this assumption begs the very question that we are trying to answer. That is, we can only be confident in claiming that voluntary actions always reflect individuals’ preferences if we are also willing to accept that individuals never act in ways that violate this link – i.e., that they never experience weakness of will. Consequently, claims that voluntary actions must
reflect preferences cannot ground arguments against weakness of will because in order to be effective, they must assume the very result they are trying to prove.\textsuperscript{71}

Having established that weakness of will is possible, the next step is determining how to approach the phenomenon through the normative framework we have developed thus far. For dealing with situations in which the two types of want diverge, it will be helpful to introduce some new terminology. In particular, we can distinguish a decision-maker’s \textit{inclination}, the option that she feels like choosing and the option that she will choose if given a free choice, from her \textit{preference}, the option she judges to be most consistent with her priorities. In terms of the previous example, I prefer to walk home, but I am inclined to take the bus.

How should we interpret rationality in light of this possible divergence between preference and inclination? What is the rational course of action for decision-makers who find themselves in a situation in which their preference does not match their inclination? In such cases, we should continue to interpret a decision-maker’s rational choice as the option most consistent with her preferences – now understood to reflect a considered judgment about priority-option consistency, as opposed to a simple inclination. The decision to conceptualize rationality in terms of preferences rather than inclination is not inevitable, but follows from a judgment that the standards for evaluating the quality of a decision should be tied to a decision-maker’s conscious assessment of her values rather than to whichever course of action she happens to feel like at the moment of decision.

\textsuperscript{71} Of course, it might still be true that revealed actions are the most consistent indicator of peoples’ true preferences. However, because this claim is ultimately empirical, it leaves open the possibility that voluntary decisions may diverge from true preferences from time to time.
More importantly, defining rationality in terms of inclination would render the concept of rationality useless for normative purposes. It follows from our definition of inclination that whenever individuals make a voluntary decision, they are necessarily selecting the option most consistent with their inclination. Consequently, when rationality is defined with inclination as its standard, referring to a decision as rational means no more than that the decision was voluntary. Unless we are willing to accept a normative framework that treats all voluntary decisions as being of equal quality, it is necessary to use preference rather than inclination as the basis of rationality. From this, it follows that acting on one’s inclination is rational only to the extent that inclination coincides with preference.72

Although we have shown that it is theoretically possible for preferences and inclination to diverge, it seems that weakness of will is the exception, not the rule. The judgment that a particular option is most consistent with one’s preferences generally bolsters the decision-maker inclination to select that option over other available choices. For example, when I judge that eggplant parmesan is the item on a menu most consistent with my priorities, I will generally be inclined to make that choice when it comes time to order. In general, the greater the utility of a particular option, the more inclined one will be to select it.73

On the other hand, one’s inclination can be affected by factors other than judgments about priority consistency. To illustrate, suppose that I walk into a fast

---

72 This conclusion also indicts an approach that is formally equivalent to defining rationality in terms of inclinations, which is to define utility functions as models of what decision-makers will actually choose to do. Such an approach would imply taking the bus must have a greater utility than walking home.

73 This point will be important in Chapter 4 because self control strategies can affect inclination indirectly by altering an option’s utility.
food restaurant and am confronted with a choice between ordering a salad and a hamburger. After considering these options, I decide that the salad is most consistent with my priorities, but as I near the front of the line, a whiff of hamburger makes me want to order the hamburger, and so I do. Because voluntary choice coincides with inclination by definition, we know that my final inclination was to order the hamburger. Whether this decision is rational or irrational depends upon why it was that my inclination departed from my original preference for ordering the salad. One possibility is that the whiff reminded me of how much I like hamburgers and altered my beliefs about which option would be most consistent with my priorities. In this case, my preference itself has shifted, and because my new inclination matched my new preference, the decision to order the hamburger was rational. However, a change in judgment is not the only reason that my inclination could have shifted from the salad to the hamburger. The smell of the hamburger may have been particularly tempting; smelling it could have made me more inclined to choose the hamburger without simultaneous convincing me that doing so would be most consistent with my priorities. If this was the case, my new inclination did not match my preference, and consequently, the decision would have been irrational.

Thus far, I have argued that particular features of a situation may cause a decision-maker’s inclination to diverge from her considered judgment. That such forces can influence behavior has been noted by others, among them Loewenstein (1996), who refers to these forces as visceral influences. In addition to adopting this

---

74 A third possibility is that my beliefs about the situation may have changed irrationally. I may have expected to enjoy the hamburger more without having a rational reason for thinking so. If this was the case, then my final decision is an example of epistemic-rather than choice-irrationality.
terminology, I will refer to an option as tempting when its presence exerts an attractive visceral influence on a decision-maker’s inclination.\footnote{By attractive, I mean that the option’s visceral influence affects the decision-maker so that she is more inclined to pick it, all else equal.} A tempting situation occurs when one of the options under consideration is tempting in this way.

The existence of visceral influences implies that there is some psychological element present in decision-making that is not fully captured by a model that explains all behavior by reference to a decision-maker’s beliefs, preferences, and constraints. In other words, some factor plays a role in shaping people’s voluntary decisions apart from conscious consideration of priorities and the perceived consequences of the available options.

We can gain further insights into the role of visceral influences in decision-making by drawing on the work of Anthony Damasio. Damasio (1994) studied patients with damage to their prefrontal cortex, the area of the brain most closely associated with decision-making. Many of these patients were able to perform the basic tasks that decision-making requires. They could identify what options were available to them, predict the likely consequences of choosing those options, and even determine the consistency of those consequences with their various goals. However, when it came to actually making a decision, these patients had a hard time selecting a final choice. What seemed to be missing in the patients was the presence of an emotional component that accompanied decision-making in uninjured people.\footnote{Anthony Damasio, Descartes’ Error (New York: G. P. Putnam’s Sons, 1994), 50.}

Based on Damasio’s work and a number of related studies, some psychologists have proposed a model of decision-making in which emotions play a
central role. Known as the somatic marker hypothesis, this theory holds that when decision-makers assess the options available to them, they typically rely on the presence of an emotional response to come to a decision about which option is best. According to this model, people make decisions by considering their available options and selecting the one that produces the best emotional response.

The somatic marker hypothesis is useful for our purposes because it highlights the role of non-cognitive factors – e.g., emotions – in the decision-making process. As people consider their various options, each option triggers an emotional response. The intensity and direction of those responses generally correspond to the consistency of an option with people’s priorities, but other factors can influence what emotion a particular option triggers. In particular, Loewenstein highlights the degree to which fundamental drives such as hunger and sexual attraction can affect the emotional valence of particular options.\textsuperscript{77} Similarly, short-term consequences tend to exert a stronger force on inclination than do consequences that are more long-term.\textsuperscript{78} This is not to say that people have priorities that attach more weight to short-term gains and losses, but rather that immediate factors tend to evoke stronger visceral influences – they are more likely to produce tempting situations. Finally, visceral influences do not need to be conscious; some feature of a situation can affect voluntary behavior without the decision-maker realizing that her inclination is being affected. This phenomenon can help explain why people may have a bad feeling about doing

\textsuperscript{77} George Loewenstein, “Out of Control: Visceral Influences on Behavior,” \textit{Organizational Behavior and Human Decision Processes} 65.3 (March 1996), 275.
\textsuperscript{78} Loewenstein, 275.
something without knowing why it is that they feel this way." In this framework, we can understand weakness of will to characterize situations in which the option associated with the best emotional response is for some reason not the option judged to be best.

Before moving on, I want to contrast my approach to other attempts to analyze weakness of will. Within economics, the most common approach has been to interpret weakness of will as a temporary change in preferences. Understood this way, willpower – a phenomenon that we will investigate in Section II – can be understood as an attempt to offset temporary preference changes in order to make behavior conform to higher order preferences. The problem with this view is that it conflates inclination with preference, and in doing so, sets up a conception of rationality in which inclinations plays a central role. As we have seen, such an approach is inadequate as the basis for a normative conception of rationality. In particular, decisions made on the basis of these “temporary preferences” are not rational in the way that decisions made on other first order preferences are rational.

Also important is the distinction between weakness of will and second-order preferences. This distinction deserves emphasis because some authors use different terminology than I do and define second-order preferences in terms of issues pertaining to weakness of will. As we have defined the term, a second order preference is a preference that decision-makers have about the content of their

---

79 This is not to say that following such intuitions is irrational. Research suggests that people may pick up on relevant features of situations unconsciously before doing so consciously. See e.g., Gladwell (2005), pp. 8-12.
preferences, not the content of their inclination. They are concerned with which option one judges to be rational whereas weakness of will occurs when one’s existing preferences do not determine the outcome of voluntary decisions. The two concepts do share a similarity, however, in that both linked to issues of control over one’s self. After all, if people could perfectly shape their first-order preferences – e.g., what makes them feel good and bad – they could ensure that their second-order preferences were always satisfied. In the same way, if people had perfect control over their inclination, they could always make their inclination conform perfectly to their preferences and they would never have the problem of giving in to temptation.

Another common approach is to investigate self control problems using multiple self models of decision-making. Rather than treating decision-makers as unified selves with unified interests, multiple self approaches assert that decision-makers are better conceptualized as a collection of heterogeneous selves vying for control of a single body. Depending on the interests of the self who happens to be in control at a particular moment, a decision-maker may be more or less prone to giving in to temptation, planning for the future, or buying honeydew melon rather than cantaloupe. Because priorities can vary widely between multiple selves, each self has an interest in imposing its own concerns and limiting the potential of other selves to subvert its goals. Weakness of will is intelligible in such models as a different self with different priorities temporarily taking control away from the self that is usually in charge. Such views are most associated with the work of Thomas Schelling and

---

80 Any preference – of first or second order – tends to promote a corresponding inclination. This is because judging that a particular option is the best way to act makes people prefer to act that way, which requires their inclination to match their preference.
have received sympathetic treatment by scholars such as Jon Elster (1986) and George Ainslie (1992, 2001).

Although unquestionably insightful for calling attention to the fact that a single decision-maker may behave in radically different ways at different times, multiple self models can be misleading when used for normative analysis. In particular, if decision-makers’ priorities are only intelligible as the collection of the interests of their component selves, how should we determine the rational course of action in situations in which the selves’ interests conflict with one another? Should the maximal option be determined on the basis of the priorities of whichever self happens to be in control in the moment or should it be determined from an aggregation of interests from all of the selves’ that constitute the decision-maker?81

The necessity of addressing these difficult questions complicates the task of constructing a normative model of decision-making that can accommodate decision-makers with multiple selves.

Even more problematic is the fact that multiple self explanations offer a misleading picture of decision-makers experiencing weakness of will. Recall that weakness of will only occurs when a decision-maker realizes that she is acting against her priorities. For this to happen, she must identify the neglected priorities as her own. When a decision-maker experiences weakness of will, it is not that her body is in the grip of a different self who has different priorities, but rather that she is unsuccessful in acting on the priorities that she identifies as her own. Furthermore,

81 Constructing an aggregate utility function from the interests of disparate selves is an analogous task to constructing an aggregate social welfare function from the interests’ of individuals. As such, it runs into analogous difficulties, most notably Arrow’s impossibility theorem.
people do not take a neutral stance between their true priorities and the temptations that give rise to weakness of will. Instead, they actively employ willpower to bring their inclination back in line with their preferences. The picture of a core self with stable priorities that emerges from this description does not fit well with Schelling’s picture of disparate selves constantly vying against each other for dominance.

The final distinction I want to draw is between weakness of will and preferences about time. Several prominent papers have explained problems of self control in terms of hyperbolic discount functions. Ted O’Donoghue and Matthew Rabin in particular have coauthored a number of influential articles pioneering this approach. In essence, they link weakness of will to the temporary preference reversals that characterize decision-makers who discount the future in hyperbolic ways. Although this approach is appealing because it explains why people would want to utilize strategies of self control without resorting to the distinction between inclination and preferences employed here, treating self control problems in terms of time preferences is susceptible to the same problems that apply to attempts to explain weakness of will in terms of any type of preference. In particular, conflating the inclinations that cause weakness of will with other types of preferences undermines the normative interpretation of what a preference is. To the extent that referring to them as “preferences” is accurate, time preferences reflect a decision-maker’s considered judgment about the best way to achieve her priorities. In contrast, the pseudo “time preferences” behind weakness of will are inconsistent with the decision-maker’s judgment. As we have seen, conflating the two concepts – and according

---

83 See Strotz (1955) for a discussion of this issue.
inclinations the status of preferences – results in a conception of rationality unsuited for normative applications. Indeed, in order to draw normative conclusions from time preference models of self control, this literature has been forced to fall back on unsatisfactory distinctions between short- and long-term preferences without establishing why one is more appropriate for the basis for a normative model than the other.

A final problem with employing time preferences to stand in for weakness of will is that time is not necessarily the relevant issue. Although self control problems are often linked to temptation for short-term benefits, it is hypothetically possible for someone to be tempted by long-term consequences rather than short-term ones. Examples of long-term temptations are few and far between because of the way that humans are evolutionarily wired to have stronger emotional responses to immediate consequences than to long-term ones.\textsuperscript{84} Despite this, formalizing the association between weakness of will and preferences for the short-term is dangerous because it implies a tighter conceptual link than actually exists.

\textbf{II. Willpower}

Now that we have developed a framework in which to approach weakness of will considerations, the next question to ask is what people should do when they find themselves in tempting situations, that is, when their inclination – “what they feel like doing” – diverges from their preferences – “what they think they should do.” Are people caught in such situations doomed to irrational behavior? Luckily, one’s

\textsuperscript{84} Loewenstein, 275.
inclination does not need to be taken as static and inflexible. When a person does not like her current inclination, she can try to change it so that it better conforms to her preferences. In common language, such efforts to adjust inclination to accord with preferences are known as willpower. Willpower can function in a number of ways, but fundamentally, it consists of a conscious effort by a decision-maker to manipulate her inclination to better match her priorities.

The most common approach through which people can control their inclination is by choosing what to focus their attention upon.85 As William James wrote, “Effort of attention is … the essential phenomenon of will.”86 By choosing to think about the adverse consequences of tempting options, we can make those costs feel more salient. Mentally emphasizing the negative effects of tempting options can associate the option with a more negative emotional response. For example, someone trying to resist the temptation to smoke a cigarette might imagine how terrible it would be to get lung cancer. Similarly, I may successfully bring about the inclination to get out of bed in the morning by considering all of the negative consequences that would result from missing class on that particular day. Whereas these examples consist of attempts to bolster the somatic markers associated with what is identified as the preferred option, another approach is to undermine the visceral influences causing inclinations to diverge from preferences in the first place. Two classic examples of this approach are for people to close their eyes and count to ten or to take a cold

85 In some cases, decision-makers may not even face a decision of whether to exert willpower. We can borrow from the literature on social norms and describe people in such situations as being in the grip of temptation. However, there are also many cases in which people recognize that what they are about to do is not what they prefer to do, and thus face a decision of whether to employ willpower to adjust their inclination.

86 William James, The Principles of Psychology (1890), 562.
shower when they feel themselves being tempted to take some unadvisable action. By mentally or physically removing themselves from the situation, decision-makers can effectively set up a wall between their inclination and the unwanted visceral influences.  

Thus far, we have been describing willpower as a method of ensuring that voluntary action conforms to preferences – a method of avoiding choice irrationality. It might seem to follow that exerting willpower is always the rational course of action. If true, this would mean that it is never rational for people to give in to temptation. However, this conclusion does not hold if we allow for there to be independent costs associated with exerting willpower. After all, many people believe that it can be good to give in to temptation from time to time, especially when the costs of doing so are not too great. Furthermore, exerting willpower is not always an enjoyable or easy task. People may prefer to avoid situations in which they will be forced to exert willpower to follow their preferences. For example, a student with an important test the next day may choose to avoid joining her friends at a bar, not because she is afraid that she will give in to temptation and drink, but because she expects that the experience will be ruined by the constant vigilance necessary to keep her inclination under control. Cases like this demonstrate that decision-makers may be reluctant to exert willpower because of the costs involved in doing so.

We can pursue this observation by considering the factors that determine the costs and benefits of exerting willpower. Willpower’s primary benefit comes from

---

87 It is important to emphasize that just because decision-makers can be successful in using willpower to overcome temptation, it is not inevitable that they will be successful in doing so. After all, people may experience weakness of will when making decisions about whether to employ willpower!
situations in which it successfully causes the decision-maker’s inclination to line up with her preferences. As such, the central issue in determining the benefit to exerting willpower is the difference between the utility of the most inclined option and the utility of the most preferred option. For example, the benefit to a student from exerting willpower to get out of bed in the morning before class will be greater on days when an important test is being given on that day. In contrast, when a tempting option is almost as preferred as the most-preferred option, the benefits from exerting willpower will be smaller. Additionally, the act of exerting willpower can be beneficial in itself. People may take satisfaction in the process of bringing their inclinations into conformity with their considered judgment and from successfully resisting temptation. In this way, some of the benefits to exerting willpower may actually derive from the employment of willpower itself.

The sources of the costs to exerting willpower are less amenable to straightforward explanation than the source of the benefits. For most people, the costs of exerting willpower vary from case to case. Individuals may be better at resisting some forms of temptation than others. For example, people who have no problem resisting sexual temptation may not be as good at following their preferences when it comes to sticking to a diet. Another way in which the costs of exerting willpower can vary is that they may depend upon how often the decision-maker has employed willpower in the past. Conventional wisdom suggests that people may be able to strengthen their reserve of willpower by using it over time. In this way, people with large reserves of willpower may find it easier to resist small temptations than people who have had less practice exerting willpower in the past. On the other
hand, decision-makers who have used willpower recently may actually find it more difficult to exert willpower again. For example, Ozdenoren, Salant, and Silverman (2005) observed that people attempting to quit smoking tend to temporarily violate their diets. Baumeister and Vohs (2003) investigated this question experimentally and found that subjects had a harder time exerting willpower when they had also been asked to do so in the recent past. Finally, there may be an opportunity cost to exerting willpower because decision-makers do not get to enjoy the psychological benefits that giving in to temptation can provide.

By acknowledging that exerting willpower can entail costs as well as benefits, we are no longer able to conclude that all uses of willpower are rational. In particular, it is irrational for decision-makers to employ willpower when the benefits of doings so are outweighed by the costs. In other words, it appears as if there are cases in which the only rational course of action is to make decisions that are not consistent with one’s preferences! Not only can giving in to temptation be rational, but in those cases making decisions on the basis of one’s preferences would actually be *irrational*.

This conclusion seems to undermine the most basic foundations of our model. After all, our entire conception of rationality is tied to selecting the most-preferred options when making decisions. The reason that our results in this section do not contradict this definition is that in tempting situations, the choice is no longer between the most-preferred and most-inclined options. Rather, the decision is better thought of as being between “exerting willpower and taking the preferred action” versus “not exerting willpower and taking the inclined action.” An analogy is the
situation in which a mugger holds you up and confronts you with the choice between your money and your life. The decision that people makes in that situation is no longer between keeping their wallet versus giving it away, but is better described as between “giving their money away and satisfying the mugger” versus “keeping their money and angering the person with a gun.” The extra consideration entailed by the presence of the mugger changes the options that one is deciding between, and can therefore change which option is rational.88 Similarly, when a decision-maker finds herself in a situation in which her inclination diverges from her preferences, this state of affairs can change the preferred course of action by linking the old best option with the extra costs of employing willpower.89

In this chapter, I have rejected the standard frameworks for modeling decision-making in favor of an approach that allows weakness of will by distinguishing between decision-makers’ inclinations and their preferences. In doing so, I have shown how our model of decision-making makes it possible for people to act in irrational ways – i.e., to exhibit choice irrationality – and have argued that this step is necessary if our theory of rationality is to retain its normative usefulness. Furthermore, the fact that inclination does not always line up with preference does not impair our ability to model behavior for descriptive purposes. Just as economists have had success in outlining the factors that can affect which option is most preferred – such as changes to income and prices – it is also possible to identify and

---

88 This is not to say that all, or even most cases of weakness of will are manifestations of a decision-maker’s rational decision to not exercise willpower. My only goal is to emphasize that such cases are possible.
89 In extreme cases, a decision-maker’s inclination may be so strongly against her preferences that no amount of willpower could bring the two back in line. In such situations, the decision-maker cannot choose the preferred option and so it is effectively no longer available part of the decision-maker’s opportunity set.
outline the visceral influences that can systematically affect inclination. Good starts in this direction can be found in the work of Elster (1998), Loewenstein (1996), and Baumeister, Heatherton, and Tice (1994).

The conclusion that weakness of will exists has fundamental methodological ramifications for modeling decision-making in the social sciences. The pervasive approach is to explain behavior by reference to decision-makers’ beliefs, preferences, and constraints, but this framework suffers from the same ambiguity that plagued Davidson’s treatment of the word “want.” In particular, it stretches preferences to cover two distinct concepts: judgment about priority consistency – what we have been calling preferences here – and inclination. Conflating the two undermines the usefulness of such models for both normative and positive applications. A model that starts from a descriptive conception of preferences – in which preferences describe what decision-makers are most inclined to choose – cannot be used for normative purposes because it treats all voluntary decisions as being of equal quality. Similarly, models that conceptualize preferences in normative terms – as judgments about consistency with the decision-maker’s priorities – will fail to describe or predict behavior in situations characterized by weakness of will. Consequently, the models that are best suited for explaining how people should behave are not be the models best suited for explaining how people do behave. Because of this disparity, the common practice of deriving normative conclusions from positive models is fundamentally misconceived.

Additionally, the existence of weakness of will raises important questions about our ethical obligations to others. People may have interests – real goals that
they care about achieving – that are not best served by the voluntary decisions that they choose to make. Examples like The Vegas Trip raise questions about how we should interact with others when we believe them to be acting irrationally. These issues also have implications for the design of policy. Legislation that attempts to influence behavior by changing the incentives associated with particular choices will have little effect if the decisions that produce that behavior are not made on a rational consideration of incentives. For example, raising the penalties for drug use may not be effective for changing the behavior of drug addicts who want to quit but lack the willpower to do so. In such cases, policymakers can achieve more success in shaping behavior by directly targeting the visceral influences that motivate irrational decisions. An example of this approach is the “cooling down” period required in many states before people are allowed to buy a handgun.

More broadly, weakness of will challenges conventional wisdom about the proper role of government in people’s lives. A classic argument against paternalistic government interference is that people can best achieve their goals when given maximum freedom to do so. The possibility of choice irrationality, however, while not directly disproving this claim, does transform the issue from a theoretical claim into an empirical one.

A surprising conclusion of this chapter is that decisions which do not seem to reflect one’s preferences may actually be rational if there are large costs to exerting the willpower necessary to adjust inclination. In other words, there may be cases in which giving in to temptation is the only rational course of action. However, the inclinations that one expects to develop in the future should not be taken as given, and
It is rational for decision-makers to take steps so that their future inclination will be aligned with their future preferences. It is to this form of self control that we turn to in Chapter 4.
In the last chapter, we saw how weakness of will can cause decision-makers to voluntarily choose options that they believe to be worse than other available options. In particular, the presence of tempting options can pull decision-makers’ inclinations out of alignment with their preferences. Although decision-makers may exert willpower to bring their inclinations back in line with their preferences, there are cases in which the effort required to do so may be prohibitive, especially when the costs of giving in to temptation are low. Thus we are unable to rule out the existence of situations in which the rational course of action is for decision-makers to act on their current inclination, even when doing so causes them to depart from their true preferences. In other words, it may sometimes be rational to give in to temptation.

This conclusion, however, is subject to an important qualification: although it may be rational for decision-makers to follow temptation *given* that their current inclination diverges from their preferences, it is often irrational to allow one’s inclination to reach that condition. For the most part, our inclinations respond to external factors in predictable ways. Alcoholics can anticipate being tempted by the availability of liquor in the same way that would-be exercisers can anticipate being tempted to skip going to the gym. In general, because decision-makers can often predict the circumstances in which they will be exposed to temptation, it is rational for them to attempt to influence and control their future behavior in order to protect
against weakness of will. Section I examines a variety of these “precommitment” strategies and investigates their effectiveness as tools of self control.

I. Strategies of Self Control

The most basic strategy for protecting against weakness of will is simply to avoid tempting situations. If a decision-maker anticipates that some aspect of a particular situation will cause her inclination to diverge from her preferences, she can prevent weakness of will by avoiding the tempting situation in the first place. For example, recovering alcoholics may choose to avoid bars because they know that the availability of alcohol will cause them to depart from their better judgment. Drug-treatment programs instruct their clients to stay away from environments and routines that they associate with drug use. Along the same lines, people trying to stick to a diet may be well advised to avoid dining at Old Country Buffet.

One problem with the pure avoidance approach outlined above is that the tempting situations may offer benefits that decision-makers would prefer to not miss out on. For example, alcoholics may be reluctant to forfeit all of the social opportunities that bars offer in order to avoid alcohol. A different self control strategy can alleviate this problem by changing the tempting situation rather than avoiding it altogether. In particular, a decision-maker might try to remove the tempting option from her opportunity set – making its selection impossible – and thereby alleviating the tempting nature of the situation. A classic example of this

---

90. It may also be rational to use self control strategies as a way of protecting against projected priority changes if the source of those changes is judged unacceptable – recall the Burmese finger puppet disease from Chapter 2. Because such situations are relatively uncommon, I focus on applying self control strategies to the type of self-conflict associated with weakness of will rather than priority change.
approach is Ulysses ordering himself tied to the mast of his ship to physically prevent succumbing to the songs of the sirens. Less mythological examples are also pervasive: a tourist visiting Las Vegas may choose to limit the amount of money she takes with her into a casino, and the hosts of a dinner party may send the leftovers home with the guests to remove the option of eating the leftovers themselves.

Although eliminating tempting choices can effectively defend against weakness of will, this strategy is not always possible. Apart from physically tying oneself to the mast of a ship, it may often be next to impossible to entirely eliminate the possibility of selecting a particular option. This is particularly true in cases where the tempting action is to pick inaction, such as in the decision of whether to stay home or go to the gym, or whether or not to procrastinate.

Rather than rendering a particular option completely unchoosable, decision-makers can attempt to manipulate the consequences attached to the available options. By attaching additional negative consequences to tempting options, decision-maker can effectively “raise the stakes” and make weakness of will less likely to occur.\textsuperscript{91} Examples of this approach are ubiquitous. People trying to quit smoking may inform their friends and families of their intentions in order raise the social costs of giving in. Similarly, consumers attempting to control their overspending can sign up to have their paychecks automatically deposited into Christmas Club savings accounts, from which they cannot make early withdrawals without paying a penalty fee. In a more direct approach, recovering alcoholics may seek to bolster their resolve by taking Disulfiram, a drug that produces an extremely unpleasant reaction when even small

\textsuperscript{91} Analogously, decision-makers can bring about the same effect by attaching additional positive consequences to the preferred option.
amounts of alcohol are consumed. Schelling describes a particularly forceful application of this strategy:

In a cocaine addiction center in Denver, patients are offered an opportunity to submit to extortion. They may write a self-incriminating letter, preferably a letter confessing their drug addiction, deposit the letter with the clinic, and submit to a randomized schedule of laboratory tests. If the laboratory finds evidence of cocaine use, the clinic sends the letter to the addressee. An example is a physician who addresses a letter to the State Board of Medical Examiners confessing that he has administered cocaine to himself in violation of the laws of Colorado and requests that his license to practice be revoked. Faced with the prospect of losing his career, livelihood, and social standing, the physician has a powerful incentive to stay clean.\footnote{Thomas Schelling, “Self-Command: A New Discipline,” in George Loewenstein, and Jon Elster (eds.), \textit{Choice Over Time} (New York: Russell Sage Foundation, 1992), 167.}

Before moving on to the next self control strategy, it is worth noting a potential objection to the incentive manipulation approach, namely that changing utility cannot be an effective tool against weakness of will because weakness of will by definition disconnects the link between preference and inclination. Reinforcing the extent to which an option is consistent with one’s preferences – raising its utility – does nothing to guarantee the option’s consistency with what matters for determining action: the individual’s inclination. There are two reasons this strategy may still be effective despite this objection. First, although inclination and preferences may depart from each other, we saw in Chapter 3 that the two tend to be at least indirectly related. Making an option more consistent with one’s preferences will, all else equal, also tend to make it more consistent with one’s inclination. In other words, options with lower utility tend to be less tempting than options with higher utility. Second, changing the utility of options can change behavior by making it more rational for
decision-makers to exert willpower. The benefits to employing willpower stem from the utility difference between the most preferred option and the most inclined option; expanding this disparity increases the benefits to using willpower and thereby makes decision-makers more likely to bring inclination back into conformity with preferences.

Whereas the incentive manipulation strategy affects inclination indirectly by changing the utility of the available options, the fourth approach is to target future inclination directly. Just as teming factors can systematically push inclination in a particular direction—away from the preferred choice—other visceral influences might pull inclination back towards the decision-maker’s priorities. By bringing other visceral influences into future choice situations, decision-makers can play an active role in shaping what inclination they develop. For instance, a number of Christian websites urge teenagers to wear “purity rings” to ensure they are reminded of their faith in tempting situations. A similar approach was the strategy used by the Argonauts to get past the Sirens. Instead of physically restraining themselves as Ulysses had done, or blocking out the song altogether by covering their ears with wax as per Ulysses’ crew, the Argonauts counted on Orpheus’s even more beautiful singing to distract them from the sirens’ songs.93

All of the self control strategies discussed thus far can be said to rely upon extra-psychic mechanisms. That is, they control future behavior by altering the circumstances in which the future decision is made. However, intra-psychic approaches to self control can also be effective, at least in some instances. One such

93 This example comes from Jon Elster, *Ulysses Unbound* (New York: Cambridge University Press, 2000), 17.
intra-psychic strategy is preference change. By altering the content of their preferences, decision-makers can defend against weakness of will by further reducing the utility of tempting options. This strategy is similar to the incentive manipulation approach discussed earlier because it works by affecting inclination indirectly and raising the benefits to employing willpower.

Another intra-psychic approach is to shape how one’s inclination will respond to particular options. This process is analogous to the process discussed in Chapter 1 by which decision-makers can change their first-order preferences in order to satisfy higher-order ones. For example, suppose that I find myself having a hard time resisting my desire to snack on chocolate, despite my judgment that snacking on carrots would be more consistent with my preferences. I could alleviate this problem by altering my tastes so that chocolate strikes me as less appetizing than carrots. In other words, by manipulating my tastes, I could ideally reach a point where carrots were more tempting to me than chocolate. Unfortunately, as we saw in our discussion of higher-order preferences, decision-makers often have at best imperfect control over what makes them happy. Altering my tastes so that I find carrots to be more tempting than chocolate would certainly solve my self control problems, but it is generally not a realistic strategy.

Before presenting the final approach to self control, I want to point out a danger with all of the strategies discussed thus far. All of the above strategies are forms of precommitment – they are methods for binding future decisions to judgments reached in a previous time period. Although precommitment can be an effective tool against weakness of will, it achieves effectiveness at the cost of
flexibility. Precommitting to a future option ensures that one does not choose a different option because of temptation, but also precludes decision-makers from diverging from a previously selected option because of improvements in judgment or to account for unforeseen circumstances. For example, depositing my money into a Christmas Club account may help curb my overspending, but if I am required to withdraw money to pay for an unanticipated hospital bill, I will be forced to pay the penalty fee.  

Similarly, business owners may regret having checked themselves into a drug rehabilitation program if their incapacitation prevents them from responding to a sudden crisis. To summarize, the binding nature of precommitment strategies that make them effective for combating weakness of will also makes decision-makers vulnerable to a range of risks due to lost flexibility. In the next section, we will consider a self control strategy that may avoid this pitfall.

II. Sticking to a Resolution

The final self control strategy discussed in this chapter is for decision-makers to resolve to follow a plan of action. Because resolutions of this type prescribe particular actions for future decisions, decision-makers who stick to them can combat self control problems by following the pre-selected plan of action. This strategy of self control offers two benefits over the other approaches discussed thus far. First, it is available in a wide range of situations. Because of its intra-psychic nature, its success is not contingent on one’s ability to manipulate external circumstances. The second advantage to resolutions is that they provide more flexibility than

---

94 Or worse: my savings, less the penalty fee, may now be insufficient to cover an important unforeseen expense.
precommitment strategies because their motivational force is contingent upon the decision-maker valuing the goal that the resolution promotes. If I resolve to quit smoking, but subsequently change my mind and decide that smoking is in fact a worthwhile behavior, I am not bound by my original commitment.\footnote{This flexibility means that resolutions are ill-suited for dealing with the other form of self-conflict we have considered, in which current selves believe their future priorities will change for unacceptable reasons. Future selves could always choose not to follow the resolution if they do not share its aims.}

On the other hand, this flexibility threatens to limit the effectiveness of resolutions as a tool of self control. For a plan to be successful, decision-makers must choose to act in accordance with its prescriptions in each time period. If decision-makers expect to experience weakness of will absent a resolution, it is unclear how the mere presence of the resolution could motivate them to change their mind. This section addresses this issue and argues that resolutions can be effective tools of self control, despite their flexibility.

One way that resolutions can modify behavior is by introducing additional symbolic benefits to choosing the maximal option. When people resolve to act in a certain way, they often perceive the act as a promise to themselves. As such, they are typically motivated to follow through on the resolution’s prescriptions, at least to some extent. For example, having resolved on New Year’s to quit smoking, I may resist buying cigarettes out of a desire to see my resolution fulfilled. However, the motivational force stemming from this desire is often quite limited and may not be enough to make decision-makers comply with a plan that they otherwise would have defected from. After all, lots of people who make New Year’s resolutions to quit
smoking do not actually quit. Luckily, resolutions can provide motivational force beyond this mechanism.

The economist and psychologist George Ainslie has been an especially persuasive advocate of resolutions’ ability to serve as tools of self control. Because he frames his arguments in the context of a multiple self conception of decision-makers – an approach we rejected in Chapter 3 – I recast and expand Ainslie’s argument using the framework we have employed. In essence, I argue that resolutions can link defections from a plan in one period to defections in future periods as well. As such, adopting a resolution can motivate changes in behavior by amplifying the consequences of failing to employ willpower. Because the arguments in support of this claim are both non-intuitive and fairly technical, it is easiest to present them in the context of a formal model.

Suppose that an individual believes that the pursuit of one of her goals will be impeded by weakness of will in future decisions. This may be because she has given in to weakness of will in the past, or because she expects to find herself in especially tempting situations in the future. To better achieve her priorities, she resolves to act in particular way: she adopts the plan that each time she is faced with a choice between options \(x\) and \(y\), she will choose \(x\). Let period 0 be the time that she adopts this resolution, so that in each subsequent period, she faces the choice between conforming to the resolution (choosing \(x\)) or defecting from the resolution (choosing \(y\)). Let \(b\) represent the net benefit of complying with the resolution over defecting in any single period – the utility difference between \(x\) and \(y\). If the decision-maker

---

96 See in particular Ainslie (1992, 2001).
follows through on her resolution in a sufficient number of periods, \(^97\) the resolution will be successful – a state of affairs associated with benefit \(B\).

The benefits to succeeding in a resolution \((B)\) may come from three sources. First, \(B\) may represent the satisfaction of a goal whose success is impeded by weakness of will. For someone who resolves to quit smoking cigarettes, \(B\) represents the utility of achieving that state of affairs. Second, \(B\) can reflect symbolic benefits associated with following through on a resolution, the extent to which people care about succeeding just because they have resolved to do so. Finally, \(B\) can represent the benefits of avoiding a “second best” strategy of self control. Because successful resolutions can achieve self control while avoiding the inconveniences and loss of flexibility associated with precommitment strategies, failing to achieve a resolution might prompt a decision-maker to adopt an alternate strategy that would entail additional costs. For example, if I want to attend parties but do not want to consume alcohol, a successful resolution may be substantially more convenient than a strategy in which I avoid parties altogether.

Finally, because we are interested in the effectiveness of resolutions as tools for overcoming weakness of will, we assume that defecting is a tempting option, so that the decision-maker must exert willpower in order to bring herself to comply with the resolution. Let \(w\) be the cost of exerting sufficient willpower to comply with the resolution in any single period.

\(^97\) The number of periods that constitute success will vary based upon the particular goal being sought. For example, if one’s resolution is to lose weight, success would follow from choosing to diet in a sufficient number of periods to obtain the desired weight loss.
To illustrate the variables we have introduced thus far, suppose that a decision-maker is attempting to lose weight and is faced with the repeated choice of whether to snack or diet. Finding himself to be repeatedly choosing to snack against his better judgment, the decision-maker resolves to diet rather than snack in each period. In this example, $b$ represents the extent to which he prefers dieting to snacking in any one period, whereas $B$ represents the overall importance to him of losing the desired amount of weight. Finally, $w$ represents the cost of exerting the necessary willpower to choose dieting when he is tempted to snack.

Returning to the model, we next consider the relation between $b$ and $w$. If $b$ were greater than $w$, then the decision-maker would find the deployment of willpower to be rational in every situation. If that were the case, however, the resolution would be unnecessary for making compliance into the rational course of action. Because we are interested in testing the efficacy of the resolution, we will thus assume that $w$ is greater than $b$. This state of affairs is unfortunate because even rational decision-makers would choose to give in to temptation in every period, thereby failing to satisfy their broader goals.

Before turning to the question of how a resolution can improve this problematic state of affairs, we need to introduce one final variable. Let $E$ represent the decision-maker’s expectation of whether or not she will succeed in following through on her planned course of action and accomplishing her goal. We can

---

98 Except to the extent that raising the utility exerting willpower bolsters the decision-maker’s inclination to do so. Because the decision of whether to employ willpower may also be subject to weakness of will, increasing the utility of employing willpower – beyond the point at which its employment is rational – can make decision-makers more inclined to employ it, and thus follow the resolved course of action.
interpret $E$ as a probability judgment so that it ranges between 0 and 1, with high values representing a large degree of confidence in the resolution’s success and values near 0 representing the opposite. Thus $E = 1$ is associated with a decision-maker’s expectation of certain success and $E = 0$ represents an expectation of certain failure. When I want to emphasize the period in which the expectation is made, I will subscript it by time. Thus $E_t$ refers to the decision-maker’s degree of confidence for overall success, as evaluated in period $t$.

We are now in a position to turn to the core of the argument supporting resolutions as an effective tool of self control. Ainslie’s fundamental insight is that our expectation about the resolution’s success – the value of $E$ – depends in large part upon whether or not we have been successful in resisting temptation in the past. All else equal, decision-makers who have chosen to defect from the resolution in the past are more likely to do so again than are decision-makers who have complied with the resolution in past decisions. To incorporate this insight into our model, we can write $E_t$ as a function of all of the choices that the decision-maker has taken since the adoption of the resolution, $E_t = f(X)$, in which $X$ is the vector of these past choices.

In order to get a feel for the basic results of this model, we can begin by making a simplifying assumption about the nature of the $f$ function – the relation between past choices and future expectations. In particular, we assume that the decision-maker starts out fully confident that she will be successful in following the resolution, but that defection in a single period shatters this belief and causes $E$ to drop to 0. In other words, we are assuming that a single past defection is sufficient evidence to the decision-maker that she will be certain to fail in future periods as
well. More formally, we are defining the $f$ function such that $f(X) = 0$ if $X_i = \text{\textquotedblleft} \text{defect}\text{\textquotedblright}$ for any past choice $i$. Although this specification is clearly an oversimplification of how decision-makers form their expectations, it will be useful for deriving our model’s implications.

To determine the effectiveness of resolutions as a tool of self control, consider the choice confronting a decision-maker at some arbitrary period $t$. As noted above, the decision-maker must choose whether to exert the willpower necessary to bring herself to choose to comply with the resolution. Furthermore, we have assumed that the benefits to complying in this particular decision ($b$) are lower than the costs of exerting willpower in this instance ($w$). To see why it may still be rational for decision-makers to choose to exert willpower and conform in this situation, we need to take into account an additional consequence of defecting, the effect on $E$. By choosing to defect, the decision-maker sends information to her future self that causes her future self to conclude that the resolution will not succeed; defection in period $t$ changes $E_{t+1}$ from 1 to 0.

This informational signal is relevant because the expectation of failure becomes a self-fulfilling prophecy. To see how, we can compare the state of affairs that results when decision-makers are not confident in the resolution’s success to the state of affairs that results when the resolution is expected to succeed.

When $E_{t+1} = 0$, the decision-maker does not expect the resolution to succeed. Because she will not have to worry about affecting future values of $E$, she will make subsequent decisions about whether to exert willpower by comparing $b$ to $w$, just like the case before the resolution was adopted. Because $w$ is greater than $b$, the rational
course of action in each future period will be to defect, and consequently, the resolution will fail.\footnote{Allowing $E$ to fall to 0 might reduce future incentives to comply through another mechanism. The benefits to complying in any one case ($b$) may themselves be contingent on expectations about compliance in future cases. For example, if I expect to gorge myself on a huge meal later, I will have less motivation for sticking to my diet now.} Is this result inevitable, or can it be avoided by a decision to comply in period $t$?

Because complying with the resolution in period $t$ maintains $E_{t+1}$ at 1, the decision faced in period $t+1$ after choosing to comply is identical to the decision faced in period $t$. Subsequently, a decision to defect in period $t+1$ dooms the resolution to failure, whereas a decision to comply maintains $E_{t+2} = 1$ and forestalls the same decision into period $t+2$. If the decision-maker avoids failure by choosing to comply for a sufficient number of periods, the resolution will have brought about its desired behavior. Consequently, it appears that the benefit to maintaining $E = 1$ in period $t$ depends upon whether $E$ will be kept at 1 in period $t+2$, which in turn depends upon whether $E$ will be kept at 1 in period $t+3$, and so on. Luckily, we can resolve this cost-benefit analysis for the decision in period $t$ without falling into an infinite regression. Because $E_t = 1$, we know that the decision-maker in period $t$ expects the resolution to succeed. Because choosing to defect in any future period ensures that the resolution will fail, $E_t = 1$ implies that the decision-maker expects to comply with the resolution in all future periods. Therefore, it follows that choosing to comply in period $t$ can eventually bring about the resolution’s success. To summarize, by maintaining $E$ at 1, the decision-maker preserves the informational effect as a compelling motivational influence in future periods, and in doing so repeatedly, brings about the resolution’s satisfaction.
The disparate consequences of complying versus defecting provide rational decision-makers with a new incentive to take into account on the side of complying with a previously adopted resolution. In addition to the other benefits of complying in period $t$ (represented by $b$), complying in a single period can make the difference between the resolution’s overall success and failure. Thus the relevant consideration for any one of the repeated choices is not $b$ versus $w$, but $(B+b)$ versus $w$. As long as decision-makers consider the overall success of the resolution to be more important than the costs of exerting willpower in a single instance, $(B+b) > w$, they will find it rational to comply with the resolution each time they face the choice. Thus the adoption of a resolution can manipulate incentives so as to augment the appeal of exercising willpower.

To illustrate how resolutions can be effective, we can examine how such a strategy would help solve a problem that has received some attention in the literature:

**The Pianist’s Dilemma**

A pianist performs nightly at a club, prior to which he eats dinner with a friend who fancies good wines. Each night the pianist is tempted to drink with dinner. He knows, however, that drinking wine will impair his evening performance, and thus prior to dinner, he prefers that he refuse the wine. Nevertheless the pianist invariably finds himself tempted to sacrifice his performance for the sake of a good Cabernet.\(^{100}\)

One way the pianist could avoid giving in to this temptation would be by removing the option of drinking wine from his opportunity set by avoiding going to dinner in the first place. Such an approach would be inconvenient, however, because

it would force him to miss out on an enjoyable dinner with his friend. The best
course of action for him would be if he could arrange to go to dinner and not have the
option of ordering wine, but such a strategy would probably be infeasible. What then
can he do? Suppose he discovers that the manager of the club has noticed the poor
quality of his performances and is considering finding someone else to replace him.
If the pianist’s strongest objective is to keep his job, does he have any option other
than skipping dinner with his friend?

Our pianist would be well-served by adopting a resolution. By resolving to
not drink wine at dinner, he can amplify the costs of giving in to temptation. In
particular, choosing to violate the resolution on one night would send a signal to his
future self that the resolution will not be successful at combating weakness of will on
future occasions, and consequently, he will have no way to keep his job other than
adopting the precommitment strategy of skipping dinner altogether. Thus, on any one
night, the desire to attend future dinners can be tapped as motivation against ordering
wine. In this way, resolutions can be used to achieve an outcome – attending dinner
and performing well – that would otherwise be unavailable.

Our simple model demonstrates how resolutions can be an effective tool
against weakness of will by making it rational to exert willpower. Like the incentive
manipulation approach described in Section I, resolutions aim to control future
behavior by “raising the stakes” of giving in to temptation. As Ainslie puts it,
“Personal rules are a recursive mechanism; they continually take their own pulse, and
if they feel it falter, that very fact will cause further faltering” (88). In other words,
the expectation of the resolution’s success becomes a self-fulfilling prophecy; when
decision-makers expect a resolution to be successful, it becomes rational to comply with it in order to preserve its motivational force for future periods. In contrast, when resolutions are expected to fail, people have no reason to consider the informational impact of current decisions on future expectations and so this motivational force is absent. The sensitivity of $E$ to past behavior means that each decision can potentially sabotage the entire resolution, and so individual acts of compliance take on greater significance.

At this point, however, it is plausible to object that our model’s results are undermined by the unrealistic nature of the assumptions that inform them. In particular, $E$ might not fall directly to 0 after a single defection – it probably requires more than a single defection for most people to lose all confidence in themselves. However, the model’s basic results are resilient to re-specifying the expectation-forming function, $f(X)$, along these lines.

Consider the case in which a single defection decreases $E$ without moving it all of the way to 0. This modification lessens the consequences of the decision made in a single period about whether to defect. When $E$ is less than 1, choosing to comply is not associated with the full benefit of the resolution’s success because the decision-maker is less confident that the resolution will succeed even if she complies today. Similarly, because defecting on a single occasion may not reduce $E$ sufficiently to doom the resolution by guaranteeing defection in future periods, the costs of choosing to defect in a particular period are also lower. Although the new model lessens the benefits of complying and reduces the costs of defecting, compliance with the resolution may still be rational. In particular, as $E$ gets lower, the benefits to
complying diminish further. If they fall below a critical value, the resolution’s motivational force becomes too weak to overcome the difference between $b$ and $w$. This may be the case for an addict who has all but given up on following through with his resolution to quit his addiction. His concern with sending a signal to future versions of himself about his resolve is mitigated by the fact that he expects himself to almost certainly fail even if he complies in any one instance. If a decision-maker attempting to follow a resolution senses that his value of $E$ is nearing this critical threshold, then a single decision acquires additional salience; complying with the resolution will be even more important to avoid pushing $E$ over the brink. Therefore, even with this more realistic specification of the model, resolutions can still be an effective tool for overcoming problems of self control.

Two features of this expanded model deserve emphasis. First, not all decisions affect expectations equally. Past decisions are only important predictors of future behavior to the extent that the situations in which the past decision was made is sufficiently similar to the circumstances that will characterize future decision-making. If decision-makers are trying to determine whether or not they will be able to stick to a resolution in the future, they will be justified in disregarding past choices that were made in circumstances that are not expected to arise again in the future. When considering the costs of defecting in a present decision, it can therefore be rational to take extenuating circumstances into account. For example, when I am trying to decide whether to stick to my diet on Thanksgiving, I can rest assured that a decision
to defect then will have limited relevance for forming future expectations about how I will make decisions at normal meals.\textsuperscript{101}

A second point of emphasis is that many of the model’s important results are sensitive to the nature of the $f(X)$ function – the relation between past choices and future expectations. In the simple model – in which a single defection shatters confidence – resolutions can make defection irrational in any instance. At the opposite extreme, consider what may be called the naively optimistic specification, in which individuals’ expectations remain at 1 regardless of past decisions. When expectations are highly resilient to contrary evidence, the informational impact of resolutions is weaker, and consequently, their effectiveness as a strategy of self control will be limited.

This section has argued that resolutions can be effective tools for self control, despite the fact that their motivational force is entirely intra-psychic. Together with their wide availability and the fact that they allow a greater degree of flexibility than precommitment techniques, it might seem as if they are the perfect strategy to combat weakness of will. Although the arguments in this section do illustrate their potential to be effective, that effectiveness is subject to several qualifications.

\textsuperscript{101} Additionally, several small epistemic irrationalities can have large implications on whether people act rationally in following the resolutions they make. Because it is rational for decision-makers to take into account the degree to which a present decision will influence future expectations, it can be rational to defect from a resolution in the presence of extenuating circumstances. However, as authors such as Akerlof (1991) have noted, people tend to systematically overestimate the importance of present features of situations. Circumstances that seem exceptional when they are present may seem less exceptional when viewed in retrospect. Thus a danger with relying on resolutions is that people may tend to underestimate the extent to which their current decision will influence their future expectations, and consequently, they might defect even when the costs of doing so are too high. In this way, resolutions can fail even when employed by mostly rational decision-makers. One way of circumventing this source of bias would be to adopt a decision-making strategy of always complying, regardless of extenuating circumstances. Such strategies will be discussed in Chapter 5.
First, resolutions are most effective when a large goal is contingent upon repeated decisions over time. For example, the goals of quitting smoking or getting into shape can be better aided by a resolution than goals that hinge on a single decision. For example, making a resolution would probably not be an effective way to ensure that I do not give in to my temptation to go out partying the night before an important exam. Even if I anticipate the temptation, making a resolution will not be an effective way to deal with it because the benefit of studying is linked to a one-time decision, and thus the informational impact of a defection will be non-existent. In cases like this, the value of $B$ is low relative to $b$ and $w$, and is thus unlikely to significantly alter the decision-maker’s incentives.

Another important qualification is that the decision-maker must initially expect the resolution to be successful. If $E$ starts at 0, or below the critical value, then decision-makers will have no reason to consider the informational impact of their decisions and the resolution will not contribute any motivational power. This consideration helps answer an objection to the resolution approach that if a resolution fails, the decision-maker can just undertake a new resolution. Although this is possible, it is likely that past failures to stick to a resolution will undermine the decision-maker’s initial confidence about the new resolution’s prospects of success, thus hampering its effectiveness.

Finally, the fact that resolutions can make it rational for decision-makers to exert willpower does not imply that decision-makers actually will exert willpower after adopting a resolution. As we have seen, the decision of whether to employ willpower is as susceptible to weakness of will as are other types of decisions. Even
when a resolution successfully motivates a decision-maker to prefer to exercise willpower, the question of whether she will actually choose to do so depends upon her inclination for that decision. Despite this, resolutions can still be an effective tool for shaping behavior to the extent that they can shape inclination indirectly through preferences. As we saw with the pianist, the knowledge that giving in to temptation on one night would force him to miss future dinners made him less inclined to order wine. In general, resolutions that amplify the benefits of resisting temptation also tend to make people more inclined to exercise willpower.

Thus far, we have focused on how self control strategies can be applied to situations of self-conflict, in particular when decision-makers’ goals are inconsistent with their expected future decisions. In addition to being useful for overcoming weakness of will, many of the same self control and precommitment strategies examined here can also be usefully employed in strategic interactions with other people. It is to these cases that we turn in Section III.

III. Credibility and Precommitment

This section is concerned with the strategic employment of self control – that is, the employment of self control strategies in interactions with other people. To see why such strategies may be useful in strategic interactions, consider the following scenario from Derek Parfit’s *Reasons and Persons*:

**The Desert Breakdown**

Suppose that I am driving at midnight through some desert. My car breaks down. You are a stranger, and the only other driver in this desert. I manage to stop you, and I offer you a great reward if you drive me to my home. I cannot pay you now, but I promise to do so when we reach my home. Suppose next that I am transparent, unable to deceive others. I cannot lie convincingly. Either a blush, or my tone
of voice, always gives me away. Suppose, finally, that I know myself to [always be rational]. If you drive me to my home, it would be worse for me [to] pay you the promised reward. Given my inability to lie convincingly, you know this too. You do not believe my promise. I am stranded in the desert throughout the night.102

The hero of this example (let’s call him Derek) cannot make a credible promise. Although the ride home is more important to him than keeping the promised reward, the structure of the situation, combined with his inability to deceive, precludes him from achieving this mutually beneficial trade.103 Before considering what steps people can take to bolster the credibility of their promises, we can consider an analogous situation in which the key issue is achieving credible threats.

**The Monopolist Deterrence Game**

Suppose that a monopolist learns that a new firm is considering entering into its market. Because the new competition will reduce the monopolist’s profit, the monopolist announces that if the new firm enters the market, it will cut prices to a level that, while quite harmful to the monopolist, will be even more harmful to the new entrant. Because the new prices would be more harmful to the new firm than staying out of the market altogether, the monopolist is confident that the threat will preserve its market intact.

However, after considering the monopolist’s threat, the new firm realizes that if it were to enter the market, it would not be in the monopolist’s best interest to follow through on the threatened course of action. Knowing that the monopolist has a reputation for strictly maximizing its profit, the new firm is confident that the monopolist will not actually carry through on its threat, and decides to enter the market.

Like the desert breakdown example, the monopolist’s threat failed because it was not perceived as credible. In both examples, the main character’s attempt to

102 Parfit, Reasons and Persons, 7.
103 Of course, the credibility problem may not arise in all cases of this type. Many people care more about keeping their promises than achieving a slight monetary gain. But for the purposes of the example we can assume this is not the case for Derek.
strategically influence the other party’s behavior was constrained by the other party’s calculation of the main character’s incentives. The crucial feature of these situations is the temporal separateness of the costs and benefits. Both situations involve two actions, one that is beneficial for the main character and one that is harmful. Because the beneficial action precedes the harmful action, the beneficial action cannot be made contingent on fulfillment of the harmful action. When the time comes to undertake the harmful action, the main character lacks the incentives to make it rational for him to do so. As a result, the other party will not expect the harmful action to be undertaken at all and will therefore treat the announced course of action as un-credible. Because the stranger does not believe Derek will pay him the reward after arriving home, he has no reason to give Derek a ride now. Similarly, because the new firm does not believe that the monopolist will ever actually cut its prices, it has no reason to avoid entry. We can illustrate these considerations by looking at graphical representations of the two examples (Figures 1 and 2 below).

The self control strategies discussed in the first section of this chapter are potential solutions to this type of strategic credibility problem. To see why, it is important to observe that the problem in these situations is not that the decision-makers do not want to make the transaction in question. The decision-makers recognize that they would be better off making the transaction – accepting both the

**Figure 1**

**Figure 2**
beneficial and harmful actions – than doing nothing at all. The problem comes from the fact that when it comes time to take the harmful action, doing so is not rational because a more preferred option is also available. Consequently, self control strategies can be effective in one of two ways. They can make the defecting option unavailable in the second period or they can change future circumstances so that the defecting option becomes less attractive. In other words, self control strategies can succeed by making it no longer rational to defect in the later period. For example, if Derek could somehow eliminate the option of not paying once he arrived home, then the stranger would have no reason to doubt the credibility of his promise. Similarly, the monopolist could commit to a business strategy that lessened the costs to it of lowering prices. With this change in incentives, the threatened action becomes the rational choice given the new firm’s decision to enter, and consequently, the threat becomes credible.
There are numerous examples of such precommitment strategies being employed as a means of achieving credibility. For example, Fearon (1994) discusses audience effects in international crises, arguing that heads of state may employ hawkish rhetoric to raise the political costs of backing down, thereby making their own threats more credible. Similarly, programming a country’s missile deployment system to automatically and massively respond to any attack would enhance the credibility of that country’s deterrence policy by removing any doubts about whether its leaders would judge such an escalation to be in their best interest. Perhaps the most important mechanism for making promises credible is the legal system. By allowing individuals to enter into binding contracts, the legal system manipulates the incentives attached to defecting so that choosing to follow an agreement is almost always the rational option.

In addition to assisting individuals seeking to make credible threats, precommitment techniques can also be utilized as a means of preempting threats from being made in the first place. Suppose that I anticipate being threatened to perform an action that, absent the threat, I would prefer not to do. I can employ a precommitment strategy to undermine the other party’s benefit from threatening me in the first place. One approach is removing the undesired action from my opportunity set. An example of this strategy are the signs in front of many convenience stores warning that “Cashiers do not know combination to safe.” Another approach is to manipulate my incentives so that performing the undesired action would be worse for me than the anticipated threat. Mafia organizations exploit
this insight when they publicly warn their members that the revenge for informing would be much worse than any punishment the police could enact.

Finally, just because a self control technique is successful in altering the rationality of a future choice does not guarantee that it will be successful in providing a strategic benefit. In interpersonal interactions, what matters is not whether future defection actually is rational, but whether the other party perceives it to be rational. Consequently, intra-psychic forms of precommitment are less likely to be effective than approaches that are externally verifiable. In a world without full transparency, measures such as adopting a resolution are unlikely to inspire confidence. Similarly, the new firm is unlikely to be swayed by the monopolist’s announcement that it has changed its preferences so that it now finds it rational to lose money by following through on its threat. However, although the importance of perception implies that some self control techniques will be ineffective in generating credibility, the flip side of this coin is that the other party may not recognize there to be a credibility problem in the first place. The centrality of perception follows from the fact that people’s incentives and intentions are not as perfectly transparent as Parfit assumes in his example. This objection, however, should not be read as undermining the importance of situations in which credibility is an issue. After all, interacting parties generally have some knowledge of the other party’s intentions, even if this knowledge is imperfect. Thus even if self control strategies are not necessary in every situation where one’s true incentives should imply a credibility gap, they are certainly necessary in many of them. Furthermore, it is likely that for every case in which credibility is not an issue when it should be, there will also be a case in which
credibility is an issue when it does not need to be. In other words, there may well be instances in which an announced course of action is credible, but the other party does not perceive it as such. Precommitment strategies may prove useful for convincing skeptical parties that decision-makers are committed to following through on their agreements, even when they would actually follow through regardless.

This section has demonstrated that the usefulness of self control strategies extends beyond situations in which decision-makers experience self-conflict. Indeed, a concern for making credible threats and promises can motivate decision-makers to exert control over their future behavior. To the extent that they find themselves in strategic interactions in which credibility is an issue, even perfectly rational decision-makers can benefit by implementing measures of self control.

Throughout this chapter, we have understood strategies of self control to be concerned with ensuring that decisions are made on the basis of one’s true priorities rather than on either inclinations or future priorities stemming from unacceptable changes. In the next chapter, we turn our attention to a different type of strategy in which people adopt an approach to decision-making whose success depends upon not seeking consistency with their true priorities. That such an approach can be rational is the subject of Chapter 5.
Once upon a time there was a robot, named R1 by its creators. Its only task was to fend for itself. One day its designers arranged for it to learn that its spare battery, its precious energy supply, was locked in a room with a time bomb set to go off soon. R1 located the room, and the key to the door, and formulated a plan to rescue its battery. There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called PULLOUT(WAGON,ROOM) would result in the battery being removed from the room. Straightaway it acted, and did succeed in getting the battery out of the room before the bomb went off. Unfortunately, however, the bomb was also on the wagon. R1 knew that the bomb was on the wagon in the room, but didn’t realize that pulling the wagon would bring the bomb out along with the battery. Poor R1 had missed that obvious implication of its planned act.

Back to the drawing board. “The solution is obvious,” said the designers. “Our next robot must be made to recognize not just the intended implications of its acts, but also the implications about their side-effects, by deducing these implications from the descriptions it uses in formulating its plans.” They called their next model the robot-deducer, R1D1. They placed R1D1 in much the same predicament that R1 had succumbed to, and as it too hit upon the idea of PULLOUT(WAGON,ROOM), it began, as designed, to consider the implications of such a course of action. It had just finished deducing that pulling the wagon out of the room would not change the color of the room’s walls, and was embarking on a proof of the further implication that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon - when the bomb went off.104

With this story, Daniel Dennett illustrates the importance of a decision-making strategy that is neither too cursory – and risks missing out on important factors – nor that involves too much cognitive analysis – whereby the process of deciding defeats the decision-maker’s broader goals. One of the chief complaints leveled against rationality is not only that is unrealistic, but that it is unrealistic for a

reason. In the real world, people do not make all, or even most of their decisions by methodically comparing the consequences of every option that is available to them. Like the robot in Dennett’s story, people who made their decisions in this way would almost certainly fail to achieve their broader goals – the very priorities that the decisions were intended to advance. In this chapter, I consider why it may be desirable, even rational, for people to make their decisions using approaches other than the optimizing method traditionally associated with rationality. I also examine how such a strategy may strengthen the elusive link between rationality and morality introduced at the beginning of this thesis.

I. Non-Optimizing Dispositions

In one sense, rationality concerns itself with the outcomes of decisions by focusing on the utility of the option that gets selected. However, there is a distinct sense in which rationality is often associated with a particular process of decision-making. Rational decision-makers are often understood to be optimizers who make their decisions by considering the full range of available options and then determining which of those options is most consistent with their preferences. Although optimizing in this way is often an effective method for identifying the maximal option, there are also cases in which rationality does not necessitate, and may even be inconsistent with, an optimizing approach to decision-making. In other words, individuals may sometimes find it rational to make their decisions using approaches other than the optimizing process traditionally associated with rationality. These alternative approaches to
decision-making, or “non-optimizing dispositions” as I call them here, are the focus of this chapter.105

The first case in which decision-makers may find it rational to employ a non-optimizing disposition is when there are costs associated with the optimizing process. The cognitive processes associated with optimizing require mental effort and attention that could be spent on more pleasurable tasks. Evidence from recent studies suggests that these costs can be substantial. For example, Iyengar and Lepper (2000) tested whether shoppers in a supermarket responded differently to small and large displays of jam flavors. They observed that although shoppers showed more interest in the larger display of jams with 24 flavors, only 3 percent chose to make purchases. In contrast, of the shoppers who stopped at the smaller display with 6 flavors, 30 percent chose to make a purchase.106 This disparity in behavior suggests that the cognitive costs from the additional options overwhelmed the benefits that the extra options offered consumers. Similarly, Iyengar, Jiang, and Huberman (2003) found that as employers increase the number of 401(k) plans that they offer, fewer and fewer employees choose to participate in any plan. The reduction in participation suggests that employees were unwilling to engage in the procedural costs required to compare all of the newly available options, and thus avoided choosing altogether.107

105 A disposition is a process for making decisions. It is important to distinguish between resolutions – which change which option is maximal – and non-optimizing dispositions – which focus on how the decision-maker chooses an option, whether maximal or not.

106 Sheena Iyengar, and Mark Teller, “When Choice is Demotivating: Can One Desire Too Much of a Good Thing?” Journal of Personality and Social Psychology 79.6 (2000), 997. These results were not simply a reflection of the smaller group who chose to stop at the 6 jam display. In absolute terms, the absolute of shoppers who chose to buy jam from the smaller display was almost eight times that of the number who bought jam from the larger one.

107 Some observers attribute the unpopularity of the 2005 Medicare prescription drug reform to the large expansion in the number of options that it offered to seniors. This reduction in participation
In addition to requiring mental effort, optimizing may be costly because of the time it takes to identify the maximal option. Depending on the situation, these costs may be more or less significant. Time lost by deciding which movie to rent is generally less costly than time spent considering the merits of every possible option when a friend starts having a heart attack. Along the same lines, researchers have estimated that firefighters make the vast majority of their decisions under intense time pressure, often having only seconds to choose from a range of possible options. Because of the particularly large procedural costs to time spent making a decision in these circumstances, it is not surprising that firefighters avoid optimizing when making their decisions.\textsuperscript{108}

There are several common ways to deal with the procedural costs from optimizing. Perhaps the simplest is to adopt a rule of thumb to be applied whenever a decision-maker finds herself confronted with a particular type of choice. Instead of comparing all of the available options, a rule of thumb prescribes choosing a particular option whenever it is available. For example, a solid rule of thumb when choosing between several unknown restaurants is to always pick the one with the most customers inside, assuming the wait is not too long. Rules of thumb are appealing when they frequently prescribe the optimal (or near optimal) option while avoiding the procedural costs associated with the optimizing process.

Another method for avoiding the procedural costs of optimizing is for decision-makers to select the first option judged to meet a predetermined standard of occurrence despite the fact that the monetary cost of participating in the new plan was substantially lower than previous plans. See e.g., Schwartz (2005).

acceptability. To illustrate, suppose that a tourist sits down to lunch in a crowded restaurant in Rome, but having neglected to learn Italian before arriving, she finds herself unable to make sense of the imposing menu before her. Luckily, she has with her a dictionary that she can use to translate the menu items. She now faces a choice about how to proceed with her decision of what to order: she can either compare all of her options by painstakingly translating the entire menu or she can begin to translate but stop at the first option that sounds sufficiently appetizing. Unless the tourist wants to spend her entire afternoon in the restaurant, it will probably be rational for her to employ the latter method.

The strategy of decision-making illustrated in this example is known as satisficing. First proposed by the economist Herbert Simon in the 1950s, the disposition has since found an enthusiastic advocate in Barry Schwartz, a psychologist and social theorist. In numerous studies, Schwartz and others have found satisficing to be an effective method for making decisions without having to consider the full range of available options. Indeed, the available evidence suggests that satisficers are more likely than optimizers to be happy with their decisions than are optimizers. ¹⁰⁹

A third alternative to optimizing is to embrace an intuitive, “naturalistic” approach to decision-making. Rather than relying on a conscious assessment of a particular option’s consistency with one’s preferences, advocates of this approach claim that people acting on “gut instincts” can often make high-quality decisions. For example, one study asked participants to rank their favorite of five jams, and found

¹⁰⁹ Schwartz, Paradox of Choice, 85.
that subjects who were instructed to carefully reflect on their options before choosing made worse decisions than those who were not given any additional instructions.\textsuperscript{110} Especially in decisions that have been faced before, studies have found that flash judgments can be at least as effective as the more methodical and reflective approaches associated with optimizing.\textsuperscript{111}

Thus far we have focused on procedural costs as a motivation for adopting non-optimizing approaches to decision-making. Another reason that decision-makers may prefer to adopt a non-optimizing disposition is when the utilities of the available options depend upon the decision-making procedure being employed.\textsuperscript{112} After reaching a decision through optimization, decision-makers tend to evaluate the benefits of the final choice \textit{relative} to the other options that were considered.\textsuperscript{113} For example, when I think back over my decision to order a particular item off of a menu, my satisfaction with the dish is tempered by the fact that I had to pass over other appealing options. As more desirable options are considered, the baseline of quality increases and the decision-maker’s satisfaction with the final decision will be lower than if fewer options were compared. Here too, non-optimizing dispositions such as satisficing and rules of thumb avoid this problem because the available options are less explicitly compared to one another.

\textsuperscript{110} T. Wilson, and J. Schooler, “Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decision,” \textit{Journal of Personality and Social Psychology} 60.2 (1991), 181-192. Decision quality was determined by consistency with experts’ ratings of the jams. Although it is possible that subjects’ in the careful choice group had tastes that systematically differed from the experts, other studies suggest that the ilicited rankings were inferior by the subjects’ own standards. See e.g., Wilson, Lisle, and Schooler (1990).

\textsuperscript{111} See e.g., Klein (1998) and Gladwell (2005).

\textsuperscript{112} Such an effect is likely when preferences are menu-dependent, a possibility that was discussed in Chapter 1.

\textsuperscript{113} Schwartz, \textit{Paradox of Choice}, 118-137.
A subset of this case is when certain goals cannot be attained through an optimizing disposition, or when using an optimizing disposition makes their attainment more difficult to achieve. Jon Elster refers to such goals as *states that are essentially byproducts.* A classic example of such a goal is spontaneity; the very act of planning one’s decisions to achieve spontaneity precludes the attainment of that end. On a more mundane level, any insomniac can attest that the worst way to fall asleep is to *try* to fall asleep. Even happiness may be essentially a byproduct in the sense that spending all of one’s time calculating which option will bring about the most happiness is not conducive to living a happy life.

States that are essentially byproducts can motivate decision-makers to adopt a non-optimizing disposition because doing so can raise the utility associated with the available options. Whereas the alternate dispositions that we have discussed thus far have explicitly avoided the process of optimizing, states that are essentially byproducts can be achieved through optimizing, as long as the goals being optimized over are not the goals *actually* being pursued. In other words, one can achieve states that are essentially byproducts by optimizing over a different set of goals. People can succeed in falling asleep by trying to count sheep and decision-makers can achieve happiness by trying to accomplish other priorities they deem important. In general, there are situations in which a decision-maker’s true priorities are best served by adopting and optimizing over a distinct set of “operational priorities.”

The final reason that it may be rational to adopt a non-optimizing disposition is to expand the set of available opportunities. This motivation arises most frequently

---

in situations of strategic interaction. To illustrate, refer back to the desert breakdown story from last chapter. Because the stranger realized that it would not be in Derek’s best interest to follow through on his promise after receiving a ride, the stranger did not treat Derek’s offer of a reward as credible. The solution that Parfit proposes for this example is for Derek to adopt a non-optimizing disposition, specifically one in which he only optimizes when doing so will not cause him to break a previously made promise. If Derek makes his decisions by following this disposition, he will choose to pay the promised reward even after receiving the ride home. By making his promise credible, the non-optimizing disposition prompts the other party to offer Derek an option that was previously unavailable: a ride home. We will return to this proposed solution later in the chapter to investigate its viability, but for now the important feature to note is the form of Parfit’s argument: adopting the non-optimizing disposition is rational because it expands the decision-maker’s opportunity set.

Before moving on, we can formalize the arguments presented thus far. A decision-maker with utility function $\pi$ must choose from opportunity set $A$, in which $\bar{a} \in A$ is the option with the greatest utility. Let $D$ represent a particular disposition such that employing $D$ for this particular decision results in the selection of one of the available options, $a^D \in A$. As a point of reference, let $O$ be the disposition of optimizing. Assuming that optimizing is an effective approach to identifying the best option, we know that the option selected by $O$ is the maximal option, $a^O = \bar{a}$. Let $N$ be a non-optimizing disposition. Consequently, we have that the option selected by $N$ must be equally or less preferred than the option selected by $O$, $\pi(a^O) \geq \pi(a^N)$. The
question that we have been addressing in this chapter is why, given this inequality, adopting a disposition other than $O$ could ever be rational. Translating the observations discussed thus far into this formal model can clarify the mechanisms at work.

In the first case, we can highlight the procedural costs of optimizing by writing the utility of employing disposition $D$ as $\pi(D)$. Dispositions with higher procedural costs are associated with a lower level of utility. The overall utility of employing $D$ for a particular decision is the utility of the selected option plus the utility of employing the disposition, $\pi(a^o) + \pi(D)$. Thus even if the utility of the choice returned by $O$ were greater than the utility of the choice returned by $N$, the procedural costs from employing $O$ might offset that difference, $\pi(a^o) + \pi(O) < \pi(a^n) + \pi(N)$. On the other hand, this inequality will not hold in all situations. Even when there are procedural costs to optimizing, it could be that employing a non-optimizing disposition would result in such inferior choices that the procedural benefits would be overwhelmed. For example, high school seniors could save lots of mental anguish by choosing where to go to college randomly, but the benefits would probably not be worth it.

In the second case, the utilities of the options are themselves dependent on the disposition that is employed. We can formalize this relation by emphasizing the role of the disposition parameter in the decision-maker’s utility function over options, $\pi(a) = \pi(a; D), a \in A$. When utility is dependent on one’s disposition in this way, it will be rational to adopt a non-optimizing disposition if the utility of the optimal option, given the optimizing disposition, is less than the utility of the option selected
by the non-optimizing disposition, given that disposition’s employment,
\( \pi(a^O; O) < \pi(a^N; N) \). For this inequality to hold, it must be that the utility-enhancing
effect of the new disposition is large enough to offset the fact that the new disposition
may not select the option with the greatest available utility.

Finally, we also saw that the decision-maker’s opportunity set may itself be
dependent on the decision-making strategy being employed. When decision-makers
choose according to disposition \( D \), we can specify that they choose from the
opportunity set associated with that disposition, \( A^D \). In this case, the non-optimizing
disposition \( N \) is preferable to \( O \) when using \( N \) results in the selection of an option
from \( A^N \) with greater utility than the option that \( O \) would have selected from \( A^O \),
\( (a^N \in A^N) > (a^O \in A^O) \). In other words, the utility of the option selected by \( N \) from
\( A^N \) must be greater than the utility of any option that was previously available.\(^{115}\)

In all of the above situations, for a non-optimizing disposition to be rational,
the benefits of adopting it must offset the new possibility of making suboptimal
choices. However, even in cases where it is rational to adopt a non-optimizing
disposition, the possibility that the decision-maker will be led to select a suboptimal
option hints at a deeper issue: the possibility of rational irrationality, in which a
rationally adopted disposition prompts decision-makers to select options that are less
consistent with their preferences than other available options. In other words, we
have shown that there are situations in which rational decision-makers will make
irrational choices! And in cases where the adoption of a non-optimizing disposition
is rational, it would be *irrational* to select the rational option by optimizing.

\(^{115}\) Because \( a^O = \overline{a} \). 

156
Before considering how we can reconcile this result with the normative framework developed thus far, we can highlight the problem by considering a striking example of it. The following scenario was inspired by Schelling but presented by Parfit.

**Schelling’s Answer to Armed Robbery**

A man breaks into my house. He hears me calling the police but knows they will not be here for fifteen minutes. The man orders me to open the safe in which I hoard my gold and threatens that, unless he gets the gold in the next five minutes, he will start shooting my children, one by one. I realize that it would not be rational to give this man the gold. The man knows that, if he simply takes the gold, I could tell the police the make and number of his car. So there is a great risk that, if he gets the gold, he will kill me and my children before he drives away. On the other hand, if I do not give him the gold, it is very likely that he will start shooting my children to persuade me to change my mind. I am in a dilemma.

Fortunately, I also have a special drug conveniently at hand. This drug causes one to be, for a brief period, very irrational. I reach for the bottle and drink a mouthful before the man can stop me. Within a few seconds, it becomes apparent that I am crazy. Reeling about the room, I say: “Go ahead. I love my children, so please kill them!” Given the state that I am in, the man is now powerless to induce me to open the safe. Threats cannot force concessions from someone who is so irrational. The man’s only option now is to flee before the police arrive.

While I am in this state, I shall act in ways that are very irrational. There is a risk that I may harm myself or my children. But since I have no gun, this risk is small. And making myself irrational is the best way to reduce the great risk that this man will kill us all.\(^\text{116}\)

The key part of this example for our current purposes is the last paragraph, which emphasizes the possibility that the narrator (henceforth Albert) will make irrational decisions while under the influence of the drug. Suppose in particular that after the thief leaves, Albert chooses to draw all over his precious stamp collection, ruining it forever. How should we assess the rationality of this decision? On the one

---

hand, it resulted from a rationally adopted disposition. The loss of Albert’s stamps is unfortunate, but certainly preferable to the loss of his and his children’s lives. On the other hand, the decision itself was clearly inconsistent with Albert’s preferences. He had invested years of his life into collecting those stamps, and felt a strong emotional attachment to them. How can we reconcile these two perspectives? Was Albert’s decision rational?

Although Schelling’s Answer to Armed Robbery constitutes an extreme example, any non-optimizing disposition can give rise to a similar problem. For example, although satisficing might be a rational way to avoid the procedural costs of optimizing, decision-makers who adopt satisficing as their disposition run the risk of making non-maximal choices. The tourist in Italy may have stopped translating the menu just before arriving at her favorite dish. The propensity of non-optimizing dispositions to produce inferior choices raises the question of how we should characterize choices that are non-maximal but emerge from a rationally adopted disposition. In such cases, I think that the best approach is to distinguish between microscopic, situational rationality and a more macroscopic, global view of rationality. Decisions like the one that ruined Albert’s stamp collection are situationally irrational but globally rational. Stated differently, we can say that a decision-maker is behaving rationally in a broad sense even while acknowledging that her individual decisions, when viewed in isolation, are themselves irrational.

117 Scholars disagree on this question. Gauthier (1998) answers affirmatively, holding that any decision which results from a rationally adopted disposition is itself rational whereas Parfit (2001) falls closer to my view that such situations are examples of rational irrationality.
The implications of this argument for the normative analysis of decision-making are enormous. To evaluate the quality of a decision, it is not enough to consider the decision-maker’s priorities and beliefs about the available options; determining whether a decision is rational also requires evaluating the rationality of the decision-maker’s disposition – the decision-making strategy that produced the decision in question. This insight is crucial because decisions that seem irrational on their face may actually reflect rational behavior by decision-makers. And more strikingly, people who optimize in all of their decisions – and thus make decisions that are each situationally rational – may actually be irrational in a more fundamental sense.

In the first section of this chapter, we investigated and formalized the situations that can prompt rational decision-makers to adopt dispositions other than optimizing. In Section II, we will focus on a subset of those situations, the case in which non-optimizing disposition aid decision-makers in strategic interactions. After examining a credibility problem that can undermine these strategies, Section III considers an application of non-optimizing dispositions to a central question in moral theory.

II. Strategic Dispositions

Now that we have established some basic results about why rational decision-makers may choose to adopt non-optimizing dispositions, we can examine more closely the use of non-optimizing dispositions to gain a strategic advantage in interpersonal interactions. In particular, adopting an alternative decision-making strategy may
solve credibility problems that result from the expectation of future optimizing. In
the first section, we saw how adopting a particular non-optimizing disposition might
be useful because of its potential to expand the decision-maker’s opportunity set.
However, I will argue that such strategies are generally ineffective at providing
credibility because they themselves are subject to a credibility problem. To address
this question, we can examine the strategy’s viability by turning once again to the
desert breakdown example.

The solution that Parfit proposes for this example is that Derek – the person
whose car breaks down – should adopt disposition $P$, which prescribes that he make
his decisions by optimizing unless doing so will cause him to violate an earlier
promise. If he makes his decisions in this way, then Derek will expect his later self to
keep the promise and pay the reward, thereby making the promised reward credible in
the first place. Discarding $O$ and adopting $P$ is a rational move because doing so
lends credibility to his promise and allows him to get a ride home. Is Parfit’s solution
viable?

The problem with this solution is that the decision of which disposition to
employ, like any other decision, is flexible and subject to revision at later points in
time. Just as Derek will not find it in his interest to pay the promised reward upon
reaching his house, he will also find it in his interest to switch back to the optimizing
disposition in order to violate his promise when the opportunity arises. The problem
here is formally the same as it was before; the benefit to adopting disposition $P$ comes
too early. When the time comes to follow through on the promise, employing $P$ is no
longer in Derek’s interest; he can benefit by switching back to $O$ because doing so
allows him to keep the promised reward for himself. As before, there is no incentive for Derek’s future self to choose to follow through on past promises, or in this case, to continue acting in a way constrained by past commitments.¹¹⁸

Is there a way out of this dilemma? One approach is to argue that the decision to reconsider an adopted disposition is itself irrational. If true, then all that matters for determining the rationality of sticking with a previously adopted disposition is whether it was rational to adopt the disposition in the first place. Using this standard deemphasizes the temporal structure of the benefits because all that matters is whether adopting the disposition is net beneficial as a whole.

David Gauthier, whose more famous argument we will consider later, makes an argument along these lines. He claims that “It is rational to continue a course of action if the expectation associated with adopting it has proved correct, even if continuing it would not then be supported by one’s outcome-oriented reasons for acting.”¹¹⁹ In other words, Gauthier is claiming that the only relevant consideration when deciding whether to reconsider a plan is only whether one’s initial expectation – that the plan will be net beneficial – is incorrect. This approach would answer our objection to the desert breakdown example because the cost of the reward is outweighed by the benefit of receiving the ride, thus making the planned course of action net beneficial overall. Consequently, even though Derek could gain by

¹¹⁸ This credibility problem does not undermine the other non-optimizing dispositions considered in Section I because employing them is linked to contemporary and future benefits for the decision-maker. For example, the reason a decision-maker would choose to satisfice when translating a menu is that she expects to gain in the present or the future from doing so.

switching back to $O$ after receiving the ride, Gauthier believes that it would be irrational for him to consider doing so.

However, Gauthier merely asserts that his is the rational approach without explaining why it is irrational for decision-makers to reconsider their past decisions when they expect to gain from doing so. In particular, he provides no good reason why the standard of rationality for this type of decision – whether or not to reconsider an adopted disposition – should be different than the standard used when making other decisions, namely whether doing so is expected to be consistent with the decision-maker’s preferences. In the decision whether to reconsider a previously adopted disposition, the key question should concern the relative expected utilities of reconsidering the adopted disposition as opposed to not reconsidering.

What factors determine whether reconsidering a previously adopted disposition will be more consistent with one’s preferences than not reconsidering? The costs of reconsidering will generally be small, such as diverted attention and time that could better be spent elsewhere. The benefits from reconsidering arise when a decision-maker determines that a new course of action would be preferable to the old one, and then makes the switch. Consequently, the benefits from reconsidering depend upon the expected difference between the utility of the current plan and the utility of the new plan that would be adopted. In other words, the more a decision-maker suspects that she would be better served by adopting a new disposition, the greater the benefits to her reconsidering the decision to adopt the old one. Thus either acquiring new information or finding oneself in a situation with new incentives can
be enough to prompt rational agents to reconsider their previously adopted dispositions.

What does all this mean for the desert breakdown example? After adopting the “never break promises” disposition ($P$) and taken the ride home, Derek finds himself faced with the choice of whether to pay the promised reward. Confronted with this unpleasant prospect (after all, he wants to keep the reward) and suspecting that his incentives may be different now than when he originally adopted the disposition, Derek rationally decides to reconsider his employment of $P$. Derek knows that he will pay the reward if he continues to employ $P$, but he also knows that in a purely optimizing disposition, he would choose not to pay the reward. Because he has already achieved the ride, there is no disadvantage to switching, and so he decides to switch back to $O$ in order to keep the promised reward for himself. Unfortunately, both parties can anticipate this line of reasoning. Adopting $P$ does not make the initial offer any more credible and Derek gets left in the desert.

The desert breakdown example demonstrates that non-optimizing dispositions can fall victim to the same credibility problems that plague other tactics in strategic interactions. Armed with this insight, we are in a strong position to critically evaluate what may well be the most ambitious application of non-optimizing dispositions ever advanced.

**III. Morals by Agreement**

In *Morals by Agreement*, Gauthier takes on a fundamental problem of ethics by arguing that it is rational for people to behave morally. Rather than claiming that the
moral course of action is always situationally rational in any particular decision – a claim we rejected in Chapter 1 – he argues that it is rational for people to reject a purely optimizing approach to decision-making in favor of a disposition in which people only optimize when it is moral for them to do so. For Gauthier, morality has a specific meaning, which is to cooperate rather than defect when faced with a prisoner’s dilemma against other moral actors. Whereas unconstrained optimizers (O) will defect in these situations, decision-makers following a moral disposition (M) will choose to cooperate.\textsuperscript{120} His argument then is not then that M and O will agree in every situation, but that in cases where they disagree, the rational course of action is to follow M.

There are two stages to Gauthier’s argument. He starts from a social contract approach, arguing that it would be rational for decision-makers to adopt M as their disposition. Because all parties do better with mutual cooperation than with mutual defection in prisoner’s dilemma, it would be rational for everyone to adopt M if everyone else were also prepared to do so. So far, I agree with Gauthier.

The crucial question is whether an agreement in which all parties adopt M could ever be credible. In other words, having made such an agreement, would it be rational for decision-makers to act according to M when doing so conflicts with O? This issue is critical because unless it is rational for decision-makers to maintain their disposition, there will be no benefits to adopting M in the first place. After all, M-

\textsuperscript{120} Although formal constraints of this type are closely associated with deontological moral theories, a purely formal conception of morality may contradict our intuitions from time to time. For example, the other party to the Prisoner’s Dilemma may be a decision-maker with aims that we take to be morally repugnant, such as a neo-Nazi or serial killer. But in Gauthier’s conception, the moral course of action in such situations is to not defect on the other party’s trust in order to better further one’s own aims, even when the content of our own aims are more moral in a substantive sense.
followers only cooperate with others when they expect the other party to be similarly constrained by \( M \). Assuming that each individual knows the incentives of everyone else – an assumption that Gauthier makes – \( M \)-followers will only cooperate with other \( M \)-followers if they can reasonably expect the other party to stick to \( M \) after entering into the interaction.

How could it be rational for someone to stick to \( M \)? After all, from a purely optimizing perspective, it is always rational to defect in such interactions – this feature is what makes the interaction a prisoner’s dilemma in the first place. Gauthier’s argument is that the other party’s choice – whether to cooperate or defect – is not independent of one’s own disposition. That is, unless the other party sees that a decision-maker is employing \( M \), they will choose to defect. Consequently, employing \( M \) and cooperating achieves a better outcome (mutual cooperation) than employing \( O \) and defecting (mutual defection). In effect, Gauthier is arguing that the utilities of the various choices are functions of the decision-maker’s disposition. The increased utility from cooperating when using \( M \) more than offsets the lost utility from not defecting, or, put into the formal language from earlier, \( \pi(a^M; M) > \pi(a^O; O) \). As we saw in Section I, arguments of this type can succeed, at least on a formal level. However, I will argue that Gauthier’s argument is flawed. In order to see why, it will be helpful to take a short detour and consider a thought experiment proposed by Robert Nozick.

**Newcomb’s Problem**

A being in whose power to predict your choices correctly you have great confidence is going to predict your choice in the following situation: There are two boxes, B1 and B2. Box B1 contains $1,000; box B2 contains either $1,000,000 ($M) or nothing. You have a choice between two actions: (1) taking what is in both boxes; (2) taking only
what is in the second box. Furthermore, you know, and the being
knows you know, and so on, that if the being predicts you will take
what is in both boxes, he does not put the $M in the second box; if the
being predicts you will take only what is in the second box he does
put the $M in the second box. First the being makes his prediction;
then he puts the $M in the second box or not, according to his
prediction; then you make your choice.\footnote{Robert Nozick, \textit{The Nature of Rationality} (Princeton, NJ: Princeton University Press, 1993), 41.}

What is the rational choice in this situation? One line of argument supports
picking only the second box: if you choose to take what is in both boxes, the being
almost certainly would have predicted this, and you will probably end up with $1000
and nothing. Likewise, if you pick only the second box, then the being will also have
predicted that choice and would have placed the million dollars in it. On this
reasoning, it appears that the correct answer is to pick the second box only.

I believe that the above argument is wrong. The key aspect of the problem is
that the being makes its prediction and places the money in the boxes \textit{before} you
make your decision. Thus your decision cannot causally affect the amount of money
in either box (if it helps, you can imagine that the money was placed in the boxes a
week ago). With the amount of money in each box fixed, it is better for you to
choose both. Because the being will have predicted this outcome, you will probably
end up with $1000, but there are worse things in life.

Armed with this new example, let us return to Gauthier. Recall that his
argument for cooperation is that the other player’s knowledge of your disposition is
what determines his choice; he will only cooperate if he sees that you make your
decisions according to $M$. The dilemma, however, is exactly the same as in
Newcomb’s Problem. Even if you adopt $M$ right up until you make your choice, it
will always be advantageous for you to switch back to $O$ and defect at the moment of decision. The other person’s expectations are formed prior to the decision and thus cannot be affected by your choice itself. As a result, $M$ is not a credible disposition; switching away from it is always the dominant strategy. Because Gauthier assumes that people’s incentives are transparent, everyone will know that it will be in everyone else’s interests to cease being moral when faced with a prisoner’s dilemma, and consequently, no one will find it advantageous to adopt $M$ in the first place.

How can one defend Gauthier’s conclusions from this argument? Gauthier’s own approach is to argue that it is irrational for decision-makers to reconsider previously adopted dispositions. We have already seen why that objection fails when we considered it in the context of the desert breakdown example. Another line of defense is to argue that Gauthier’s transparency assumption is the source of the problem. If decision-makers’ preferences are not fully transparent, then others will only have imperfect information about their incentives and dispositions. Consequently, potential cooperators will have to rely on past actions as a source of information about how a particular decision-maker can be expected to behave in the future. This type of reputational effect, it has ben argued, can provide the decision-maker with an extra incentive to cooperate. In effect, making past behavior relevant to future opportunities transforms the one-shot interaction into a repeated prisoner’s dilemma. This change undermines our objection to Gauthier because sticking with $M$ and choosing to cooperate may now be rational.

Although introducing reputational effects does bolster the credibility of $M$, it does so at the expense of Gauthier’s broader objective. First, this approach does not
establish morality as a rational constraint on optimizing. By altering the incentives so that cooperation is the optimal course of action in most interactions, we have shown that morality and optimization often prescribe the same behavior, but we have not shown that it is rational to act morally when the two prescriptions differ. In other words, we have shown that it is often rational to act morally, but not that morality is a rationally necessary constraint on choice.

Second, even if it is true that morality and rationality tend to coincide in most situations, the above approach also suggests that there will be cases in which the two diverge. The reputational effects that lie at the heart of this argument motivate people by deterring defection out of a concern for participating in future interactions. For cooperation to be rational, decision-makers must believe that defection today will result in others refusing to cooperate with them in the future, and that the costs from this foregone cooperation will outweigh the benefits gained by defecting today. This threshold will not always be met, particularly when opportunities for future cooperation are limited or when decision-makers believe that others will be willing to cooperate with them again no matter what. Additionally, there is the well known problem that when individuals expect to repeat such interactions only a finite number of times, they may employ backwards induction and choose to defect immediately. Even more problematically, defection is only costly when information about prior defections is available to potential cooperators. As such, reputational considerations can only deter defection to the extent that past defections can be detected and warned against. This condition may not always be satisfied. For example, if two people are stranded on a desert island, one will not be deterred from defecting in a way that costs
the other his life, as long as doing so ensures that future cooperators will not find out about his defection.

If we embrace the reputational line of defense, we are left with a severely weakened version of morality, one that would not be binding for decision-makers in situations where reputational effects are weak. The resulting prescription says to act morally when doing so is convenient, but that immoral action is fine so long as one does not expect to get caught. The inadequacy of such a view of morality is clear evidence against the success of Gauthier’s project.

Has Gauthier’s claim been defeated? We have seen that introducing reputational effects is not an effective approach, but there is another, more promising tactic. Earlier in this chapter, I argued that it is rational for decision-makers to reconsider their adopted dispositions when they expect to benefit from doing so. At the same time, I rejected Gauthier’s assertion that it is only rational for decision-makers to reconsider adopted dispositions when they expect that their original decision was in error. What I now propose is that although Gauthier’s standard is inconsistent with optimizing behavior by decision-makers, it might still be rational for decision-makers to adopt Gauthier’s standard as the standard they use to decide whether to reconsider their dispositions. In other words, I am proposing that people adopt a second-order disposition to guide decision-making about sticking to other, previously adopted dispositions.

The arguments advanced throughout this chapter suggest an immediate objection to my proposed approach. Decision-makers could choose to discard this second-order disposition as easily as they can choose to switch away from their other,
first-order dispositions. Although I admit that decision-makers will have the option of discarding this new disposition, it will often be irrational for them to do so. The reason this second-order disposition can succeed is that it transforms what is essentially a one-shot decision – the choice of whether to switch dispositions – into a repeated decision. As such, the intra-psychic strategies discussed in Chapter 4 – namely adopting a resolution – may be successful here as well.

Suppose that a decision-maker resolves to not switch away from a previously adopted disposition unless he comes to believe that his original decision to adopt the disposition was in error. Specifically, let us suppose that Derek – from the desert breakdown example – has adopted disposition $P$ in which he always fulfills his promises, as well as the second-order disposition described above. Before, the credibility problem arose because it was in Derek’s interest to switch back to an optimizing disposition when it came time to pay the reward. Now, however, such a defection would violate the second-order disposition because Derek still believes that paying the reward would be better than being stuck in the desert. As such, his decision now becomes whether or not to break the resolution.

To see why it may be rational for Derek to stick to his second-order disposition here, we can turn to our discussion of resolutions from last chapter. A plausible reason for not breaking a resolution is that it can send a signal that one will be unable to follow through on the resolution in future cases. If Derek adopts $P$ just to discard it when it becomes convenient to do so, he will have less faith in his ability to follow through on this type of non-optimizing dispositions in the future. In other words, violating his resolution now will make it harder for him to credibly adopt non-
optimizing dispositions in the future, in situations with temporal structures that gives rise to credibility problems. If Derek wants to make credible threats and promises in the future, it may be rational for him to stick to his resolution – and keep $P$ – in order to allow himself to credibly adopt non-optimizing dispositions in these future situations as well. By the same logic, decision-makers who resolve to adopt this second-order disposition can overcome the credibility problem associated with Newcomb’s Problem, and may now find it rational to stick with $M$.

By introducing this resolution, and demonstrating that it can be rational for decision-makers to follow it, we have come closer to proving Gauthier’s intended result. We have shown that as long as decision-makers care about being able to credibly adopt non-optimizing dispositions in future interactions, it may well be rational for them to comply with $M$ as a constraint on pure optimizing. However, the necessity of this extra step makes Gauthier’s result contingent on decision-makers finding themselves in particular circumstances, circumstances that may be common but need not be universal. Although it will often be rational for decision-makers who have adopted $M$ to conform to it, the same will not be true for people who, for whatever reason, are not as concerned with being able to adopt credible non-optimizing dispositions in the future. Consequently, the constraint that we have shown to be rational is inadequate as a conception of morality – its motivational force is limited to people who find themselves in a particular set of circumstances. On the basis of our arguments then, $M$ is still only a hypothetical imperative, not a categorical one.
Although we have challenged Gauthier’s claim that non-optimizing dispositions can tie together rationality and morality, our discussion of the issue has highlighted the importance of accounting for dispositions in assessing questions of rationality. Furthermore, although we fell short of proving Gauthier’s intended result, our conclusions suggest a larger coincidence between rationality and morality than is commonly assumed. In particular, we have shown that it is often rational for decision-makers to choose to cooperate when faced with a prisoner’s dilemma, despite the fact that such choices appear irrational when considered in isolation.

Our exploration of rationality is nearly complete. In this chapter, we examined a variety of decision-making strategies other than optimizing and showed how adopting those strategies can be rational, even when doing so causes decision-makers to select non-maximal options. This recognition – that it may be rationally necessary to make (situationally) irrational decisions – is certainly one of our most surprising results. If noting else, it underscores the need for normative theories to take a holistic perspective when assessing the quality of a decision.
This thesis has developed a normative framework for analyzing decision-making and applied it to a variety of situations in which decision-makers experience self-conflict. Using the tools and concepts developed in Chapter 1, we looked at what rationality requires of decision-makers in situations characterized by intertemporal conflict between current and future priorities (Chapter 2) as well as situations of intratemporal conflict between priorities and inclinations (Chapter 3). In Chapter 4, we considered a range of self control techniques that decision-makers could employ to ensure consistency between their behavior and their priorities. We also saw how rational decision-makers could use these strategies to gain strategic advantages in interpersonal interactions. Finally, Chapter 5 showed how decision-makers could benefit by employing methods of decision-making other than the optimizing process traditionally associated with rationality.

A theme that emerged in the course of our discussion is that the rational way to deal with self-conflict depends on the validity of the two sides in conflict. When priorities change over time for acceptable reasons, both sides of the conflict represent the decision-maker’s authentic interests, and therefore rationality demands the exercise of prudence. In contrast, when one side of the conflict is less valid than the other, as in the cases of weakness of will and unacceptable priority changes, rationality implies that decision-makers should favor the side of the conflict that better reflects their true concerns. This result explains why decision-makers may find
it rational to adopt self control strategies in which they attempt to further one side of
the conflict against the other.

In the process of constructing our model of rationality, we have accumulated a
rich collection of tools with which to analyze decision-making. In the remainder of
this conclusion, I take up two issues that demonstrate the applicability of the concepts
we have developed. The first concerns the relation between social forces and
individual decision-making and the second challenges the conventional wisdom that
there is a tight link between freedom and well-being.

I. Decision-Making and Social Forces

Throughout this thesis, we have repeatedly emphasized the project’s normative
nature. Our model of rationality is tailored to addressing normative questions, rather
than descriptive ones. Yet many of the concepts that emerge from a normative theory
of decision-making have positive applications as well. To illustrate, we can apply the
concepts developed here to a central question in the social sciences: the influence of
social norms on individuals’ decision-making. Although this question is descriptive
rather than normative because it seeks to explain behavior rather than prescribe it, the
concepts we have developed here can provide insight into the interplay between
structure and agency in explaining this phenomenon.

A perennial and fundamental source of controversy among academics
concerns the relative importance of societal and individualistic explanations of
behavior. On one extreme of this methodological divide is the structuralist tradition
associated with Marxists and sociologists in the tradition of Emile Durkheim. At the
other end of the spectrum are economists such as Gary Becker, who attempt to
explain all social phenomena in terms of the decisions of rational agents. Although few economists believe that all human behavior derives from rational judgments, and few structuralists believe that people are entirely the playthings of social forces, the methodological divide between the two camps is quite real.

However, as many have noted, the dichotomy between structure and agency is analytically counterproductive. Social phenomena do not exist independently of the individuals whose behavior constitutes them, and individuals do not make their decisions in a vacuum, isolated from the social world. An adequately broad model of decision-making reveals the extent to which social forces can shape the choices that people make, while also acknowledging the importance of individuals’ intentional pursuit of their goals. In contrast, a methodology that focuses exclusively on structure or exclusively on agency risks overlooking important mechanisms through which each can be important for explaining behavior. The conception of decision-making developed in this thesis can help reconcile the two approaches by shedding light on how social forces can be incorporated into a model of behavior oriented around individual decision-making. To illustrate the importance of a methodology that allows for both structure- and agency-based explanations, we can consider an issue that epitomizes the divide between the two methodological camps: the role of social norms in shaping behavior.

To begin, we can follow Elster (1989) by defining a norm to be a prescription for action that is not conditional on the prescribed behavior’s expected outcome. In contrast to instrumental prescriptions of the form “Do X, if doing X will bring about
some desirable goal Y,” norms take the form “Do X” or “Don’t do X.”122 For a norm to be social, it must be shared by other people and partly sustained by their approval and disapproval.123 Whereas structuralists take social norms to be the basic motivational force behind behavior, economists typically explain them as derivative of conscious decisions; people follow social norms only to the extent that they choose to do so. In what follows, I describe a variety of mechanisms through which social norms shape individuals’ behavior. Some of these mechanisms are tied to the intentional pursuit of goals, but some are not. As such, overly stylized models of decision-making are inadequate to explain social norms’ motivational force.

Social norms can influence individuals’ behavior in a variety of ways. Some of these motivations can be accounted for through decision-based models of behavior. First, people might behave in the way prescribed by a social norm because the prescribed behavior coincides with what the decision-maker believes to be rational. For example, decision-makers may choose to conform to truth-telling norms because they believe that choosing to lie would end up being worse for them in the long-run – perhaps because they prefer telling the truth to the burden of keeping track of a web of lies. Second, decision-makers may find it rational to adopt a social norm’s prescription as their disposition. In other words, making choices based on the norm may be a rational approach to decision-making, despite the fact that the option prescribed by the norm might not be maximal in all instances. For this reason, it can be rational to follow a social norm in the same sense that it can be rational to adopt a

122 Jon Elster, The Cement of Society (New York: Cambridge University Press, 1989), 98. As we will see, this imperative need not be conscious; social norms can push behavior in the direction of the prescribed action in unconscious ways.
123 Elster, Cement, 99.
rule of thumb. Finally, decision-makers may find it rational to follow a social norm in order to avoid the sanctions that society imposes on violators.\textsuperscript{124} For example, consider the norm against men wearing women’s clothing. Males who would otherwise prefer to wear a skirt in public may choose not to do so to avoid the social shunning that breaking the norm would entail.\textsuperscript{125}

In the first two cases, rational decision-makers would choose to follow the norm-prescribed behavior even if there was no social norm; the rational decision happens to coincide with the option prescribed by the norm. In the third case, decision-makers follow the norm because it is in their best interest to do so, but it is only in their best interest because of its status as social norm. In other words, if other people did not care about the norm, decision-makers would not be motivated to follow it. All three of these motivations lend themselves to analysis with models of individual choice and demonstrate why a purely structural explanation of social norms – which ignores individuals’ decisions about how to pursue their goals decision-maker’s preferences – is insufficient. On the other hand, there are a number of mechanisms through which social norms influence decision-making that are less amenable to rational choice explanations.

Most basically, social norms play a large role in determining when people make decisions in the first place. As we discussed in Chapter 1, many actions do not

\textsuperscript{124} The question arises, however, why other members of society would choose to enforce these sanctions. After all, there are generally few sanctions imposed on those who fail to sanction norm-violators, and even fewer on those who fail to sanction them. Thus it may be necessary to invoke non-rational mechanisms when explaining the behavior of those other members of society who impose the sanctions that the decision-maker is rationally motivated to avoid. See Kaushik Basu, \textit{Prelude to Political Economy} (New York: Oxford University Press, 2000), 87-88.

\textsuperscript{125} Although other considerations may also be a factor, such as the disparate statuses associated with men’s and women’s clothing.
emerge from decisions in the sense that people undertake them without consciously considering any alternatives. Social norms can influence behavior in this way by affecting which actions people do unconsciously and which actions are only done after being decided upon. For example, members of some communities might immediately return a smile or a wave of greeting out of habit, whereas members of communities that lack this norm might do so only after some consideration.

Furthermore, when people do make decisions, social norms can affect which options get considered. For example, upon reaching the end of a taxi ride, most people do not even consider the possibility of jumping out of the car without paying the driver, no matter how much utility would be associated with this option. Moreover, social norms also influence behavior that does not emerge from decisions. For example, most people do not consciously decide how far from others to stand during polite conversation, but the preferred standards of personal space vary widely between cultures.

Third, social norms influence decision-makers’ priorities and values. People’s priorities are largely shaped by their past experiences as well as by the values of those around them. Decision-makers who are raised in cultures that subscribe to a particular norm are likely to internalize it as their own, and to make achieving that norm a priority in their own decisions. Although it has long been recognized that

---

127 Elster, Cement, 106.
preferences are endogenous in this way, there have been relatively few attempts to incorporate the phenomenon into economic models or to investigate it empirically.\textsuperscript{128}

Finally, social norms influence individual decision-making by affecting people’s inclinations for particular options. An option that produces a positive visceral response in one culture may have the opposite effect in a different society. The emotional responses that people feel when they consider a particular option are shaped by the object’s meaning to them – a meaning which is socially constituted through both language and culture. For example, norms against homosexuality are partially sustained when members of society have strongly negative visceral reactions to displays of affection between people of the same sex.

Although we cannot embark on a more thorough consideration of these issues here,\textsuperscript{129} our brief discussion of social norms illustrates how social forces shape individuals’ decision-making. Social forces influence individual behavior by affecting intentional actions – by sanctioning people who violate norms, for example – but also in more subtle ways that have traditionally been focused on more by social theorists than by economists.\textsuperscript{130} This recognition erodes the conventional dichotomy between structure and agency by acknowledging the ways in which social forces social forces filter through individuals’ decision-making to affect behavior in the

\textsuperscript{128} Bowles (1998) provides a good discussion of this issue. Further examination into this mechanism would invoke questions such as whether it can be rational for decision-makers to embrace priorities that are only persuasive for people who grow up in a particular culture.

\textsuperscript{129} For example, another mechanism is the role that social norms play in affecting the beliefs that inform individuals’ decision-making. A fuller conception of rationality that included a model of rational beliefs would be better equipped to address this issue.

\textsuperscript{130} A more ambitious project could examine whether conceptions of power in the work of such thinkers as Michel Foucault could be reconciled with the models of individual choice considered here. One stumbling block would be Foucault’s broader criticisms of methodologies based on enlightenment concepts such as rationality and agency.
aggregate. Attempting to explain a social phenomenon without considering both decision-making and social forces requires assuming that the mechanisms in the neglected category do not play a role – an assumption that will rarely be justified.

II. Choice and Well-Being

The conclusions reached in this thesis also have relevance for another fundamental question in the social sciences, the relation between choice and well-being. In traditional neoclassical models of behavior, providing decision-makers with access to additional choices can only aid in the satisfaction of their preferences. Because decision-makers are always free to not choose the additional option, so the argument goes, the added option will make a difference only if the decision-maker chooses it, in which case it must be associated with more utility than any option that was previously available.

However, a number of our results imply that there are cases in which adding more choices can make decision-makers worse off. There are several distinct mechanisms through which this effect can occur. First, adding choices can make decision-makers more susceptible to weakness of will. In particular, the addition of a tempting option will be counterproductive if decision-makers choose it over other options that are more consistent with their preferences. Ulysses would not have been better off had he been given the opportunity to be untied from the mast. Second, even when they are maximal, the presence of additional choices can make decision-makers worse off by diminishing the credibility of their threats and promises. In Chapter 4, for example, we saw that the availability of the option to renege on a promise can make it harder to secure a ride home from the desert. A third possibility is that
adding more choices can increase the procedural costs required to make a choice. As the 2005 Medicare drug reform demonstrated, the flexibility offered by new choices may be offset by the additional effort that considering these new options would entail. Finally, when preferences are menu-dependent, the utility of the maximal option may be lessened by the presence of additional options that are non-maximal. The satisfaction that one feels from the maximal option may be lessened if the new choice raises the baseline for comparison.

That additional choices can decrease the extent to which rational decision-makers satisfy their priorities drives a wedge between two otherwise compatible conceptions of well-being: a utilitarian account based on preference satisfaction and a capabilities approach in which well-being is derived from the size of a person’s opportunity set.\textsuperscript{131} Suppose that a new policy provides a decision-maker with an additional option, thus expanding the set of capabilities from which she can choose. In a capabilities sense, policies that have this effect make decision-makers unambiguously better off. However, the above discussion highlights the possibility that the new choice will actually decrease the extent to which the decision-maker brings about the satisfaction of her preferences. In this utilitarian sense, the policy has the opposite effect on the decision-maker’s well-being. The potential for a single change to simultaneously expand capabilities but decrease the satisfaction of preferences forces us to decide which conception should be used as the basis for evaluating well-being.

\textsuperscript{131} See Sen (1999) for a more in-depth discussion of the differences between these approaches.
Additionally, the potential for extra options to make people worse off (in terms of preference satisfaction) raises important questions about the proper role of government intervention into the lives of citizens. Government restrictions on citizens’ private behavior – behavior that does not substantially affect others – have been criticized on utilitarian grounds by thinkers such as Bentham and Mill. The common assumption grounding such criticisms is that citizens can best achieve their goals when given maximum freedom to do so. However, our results call this assumption into question. Because situations exist in which access to particular options makes decision-makers worse off, policies that restrict citizens’ choices have the potential to increase their well-being.132

Others have put forward ideas along these lines. Scholars such as Jon Elster (2000) and Cass Sunstein (1991) have suggested that constitutions can be understood as tools of self-restraint that citizens impose on themselves. Moreover, a number of existing government policies are most intelligible as tools of self control, such as restrictions on anticipated tax refunds being used for down payments on short-term loans.

Although such a view offers a plausible justification for government interference in private behavior, it runs into an important problem. In particular, the options that it is beneficial for the government to restrict will vary depending on the priorities of the citizen in question. Because of the wide range of priorities that citizens may have, an option whose presence is counterproductive for one person may

132 In most cases, a government policy cannot actually make a prohibited option physically impossible, but can only attach negative consequences to its undertaking. Although the two types of prohibitions are clearly distinct, it will be convenient to refer to both as restrictions on choice.
be helpful to another. And because we have seen that most priorities are at least potentially rational, it is generally impossible to determine with certainty whether the presence of an option is making a particular decision-maker worse off. As such, any policy that seeks to assist people by eliminating unhelpful options runs the risk of interfering with the rational pursuits of other citizens.\textsuperscript{133}

Furthermore, because it is likely that these restrictive policies would interfere with the pursuit of some priorities more than others – priorities whose rational pursuit entails behavior that is similar to others’ irrational behavior are likely targets – an approach to government formulated along these lines could systematically put citizens who held those priorities at a disadvantage relative to citizens with other priorities. Consequently, such policies might be inconsistent with one of Liberalism’s fundamental tenets, that political systems not discriminate between the reasonable viewpoints that citizens may have.\textsuperscript{134} The question of whether choice-restricting policies may still be justifiable in light of such concerns is an important one, and merits further consideration.

The issues addressed in this conclusion are fundamental ones, and our treatment of them has been far from comprehensive. What our brief discussion does indicate is the extent to which the tools developed in this thesis can inform the way we think about a broad range of questions in the social sciences. And what of our

\textsuperscript{133} An analogy can be made to a question of dining policy at Wesleyan. Some of my friends have remarked to me that they wished Mocon – the all-you-can-eat cafeteria on campus – only allowed patrons a single trip through the food line so that it would be easier for them to resist the temptation to eat more than one serving. However, for such a policy to be effective, it would have to affect all patrons, which would be unfortunate for students like me who enjoy having the option of making multiple trips.

\textsuperscript{134} This tension suggests a role for individual rights in limiting the reach of choice-restrictive policies. By creating a sphere of fundamental, protected options, individual rights could ensure that all citizens had access to a minimum level of tools with which to pursue their reasonable priorities.
results concerning rationality itself? After all, we began this project by noting the value of a normatively compelling account of rationality, both in its own right and because of the larger projects that rationality informs. Have we succeeded? Is our theory of rationality up to the task?

There is plenty of cause for optimism. Our model has overcome many of the obstacles that other accounts of rationality founder over. It proved successful at analyzing common forms of self-conflict and confirmed the wisdom of decision-making strategies that elude traditional approaches. In these ways, we have narrowed the gap between theory and reality, between a normative model of rationality and a common sense understanding of what good decision-making entails.

Yet our picture of rationality is still incomplete. For every type of situation that we have considered, there are countless others in which decision-makers may find themselves; and the rational course of action will not always be apparent. Although it would no doubt be illuminating to expand our model to cover all such cases, striving for absolute completeness seems like a hopeless task. Life throws more puzzling situations at us than we could ever comprehensively address in a theory of rationality, or anywhere else. A model of rationality cannot tell us how to live our lives.

But it can point us in the right direction.


